BACHELOR THESIS

Improving the number of employees at a control centre

Martijn Korver
m.c.j.korver@student.utwente.nl

Industrial Engineering & Management

SUPERVISOR PRORAIL
Saskia Wevers

SUPERVISORS UNIVERSITY OF TWENTE
Ipek Seyran Topan (1st supervisor)

Engin Topan (2nd supervisor)

UNIVERSITY OF TWENTE
7522NB Enschede
Drienerlolaan 5

PRORAIL
3527KV Utrecht
Admiraal Helfrichlaan 1

Foreword

My bachelor thesis is called: “Improving the number of employees at a control centre”. This report is the final test for the bachelor study Industrial Engineering and Management at the University of Twente. The research is performed at the OBI department of ProRail in the control centre in Utrecht. The current situation is with regard to the amount of employees is analysed and via a mathematical and simulation model improved where possible.

In this foreword I want to thank my supervisor from ProRail, Saskia Wevers. I was always able to contact you for anything, whether it was about ProRail or about the report. Furthermore I want to thank Delailah de Lima, team manager of the OBI department, and all the employees working at the OBI department. I could ask them any question and always got an helpful answer in return.

Besides, I want to thank Ipek Seyran Topan and Engin Topan for being my supervisors of the University of Twente. They gave me clear feedback and useful tips, which really helped me during this report.

Lastly I want to thank everybody else that helped me during this research.

All that remains for me to say is: enjoy reading my bachelor thesis.

Martijn Korver
Utrecht, June 2019

Management summary

Problem description

The railway system of the Netherlands is a complex system, with many different assets needed to let the trains drive. One of those assets is the electricity network. This electricity network consists of catenary, traffic lights and switches. This network has to be monitored at any time, which is done by the employees of the OBI department. Since the workload at the department can very fluctuate, this research will look into possibilities of scheduling less employees at some shifts during the week. There are multiple stations, for two of them the research will be done, first both separate, later also the combined station will be researched.

The goal of this research is to determine the minimal number of employees needed to run the department successfully. Therefore the following research question is stated: ***Is there a way of scheduling the employees more efficiently, such that less employees are needed to perform the same amount of work?***

Literature research

In the literature, models have been found which are commonly used in the workforce scheduling cases. These models are: crew scheduling, days-off heuristic and shift scheduling. By looking into these models, a pattern has been found on how to solve workforce scheduling problems.

1. First the case has to be made in detail, so the constraints have to be determined, the amount of employees, the workload etc.
2. The findings in step 1 have to be put into a model, mostly a LP model. So everything has to be put into variables. After that has been done, an optimization formula has to be derived, which will give a solution to the problem.
3. After the model is made, it has to be solved. As solving optimally takes a lot of time, the model is solved until one solution has been determined. This solution is the 'best solution' for now.
4. With the solution of step 4 as initial solution, algorithms and heuristics, like the Day-Off heuristic will be used, as well as regression on all the possible neighbour solutions. These neighbour solutions can be determined by local search.
5. After finishing step 4, all outcomes can be analysed and the best outcome will be chosen as the schedule.

In this case, the LP model can easily be solved, so step 4 and 5 are not needed. However, to be sure that the outputs of the LP model really fit the department, a simulation model will be made for the company. By keeping track of the warm-up period, run length and the number of replications, the department will be simulated.

Data analysis

For the gathered data sets, distributions will be fitted such that the data can be inserted into the simulation and mathematical model. For this analysis R is used, plotting the data and using the Kolmogorov-Smirnov tests to find the best suited distribution. For the mathematical model the 99% confidence-interval is used to insert into the model.

Mathematical model

The mathematical model is an integer linear programming model, which determines the number of employees to start their shift at a certain time. Solving the model for a week gave the minimal amount of employees needed to run the department. However, due to the high variability, the decision was made to verify the solutions of the mathematical model via a simulation model.

Simulation model

The simulation model is made to be able to verify the solution of the mathematical model. The output of the mathematical was inserted into the simulation model to determine average waiting times for the jobs. All the outputs for the different input values were analysed and resulted in the conclusions and recommendations.

Conclusions

The mathematical model showed that for the stations North 1500V and the Combination 1500V, there were some shifts for which the number of employees could be decreased. Especially the evening shift from 15:00 – 23:00 was a shift that needed little employees to be fulfilled. The night shift could not be decreased at all, since the effect on the waiting time for the sequences is big and the work really has to be finished within the hours, since the maintenance has to start.

The lowering of the number of employees during the evening shift, however, increased the utilization level of the North station drastically. With the addition of the extra manual switching handlings, the utilization level is almost 100%.

Advice

My advice for ProRail and the OBI department will combine all the station conclusion into one advice. First the actual advice will be stated, after which the arguments will be given.

- Keep the stations separated, set the number of employees working at North during the evening shift to 1, but add an extra flexible employee working other times than the normal shifts and able to work on both North and South station.

Why decreasing the number of employees working the evening shift at the North station to 1 and add a flexible employee? The reason to decrease the amount of employees in the evening shift is because the increasing waiting time of the jobs is little compared to the current solution. Furthermore, this increasing waiting time is for a very large part based on the failure assumption in the simulation model. As explained earlier, this failure assumption increases the waiting time a lot. Therefore, when looking at the mathematical model as well as the simulation model, in my opinion the company could decrease the number of employees working on the North station in the evening shift. The reason to add the extra flexible employee is because the utilisation of both stations is quite high during the morning and evening shifts. Especially the hour from 16:00-17:00. By adding an employee which works from 09:00-17:00 and can work on both stations, the peaks can be managed more easily. This means that all employees work effectively during their shifts.

The addition of the extra flexible employee, will not increase the number of employees needed per week, but will increase the flexibility of the workforce scheduled. In the most busy hours there are more employees, and in the quieter hours, less employees are scheduled. Also the work pressure on the employees might decrease, which is both beneficial for the employees as well as the company. Employees that feel less pressure often has more fun in its job.

The combination station will decrease the amount of variability in the system. Since the unknown time of the manual switch handlings, there is even more variability in the North and South stations. By combining them, the variability decreases. However, it is to the company to think of whether the amount of variability is large enough to really benefit from it. In this case, the potential benefit of combining the stations is little, therefore the decision to add a flexible worker would be more beneficial. This is a little step towards the building of the combination station.

Content

s

List of tables

List of figures

Chapter 1: Introduction

This chapter gives an introduction to the company and the problem. The research objective as well as research questions and the plan of approach will also be handled.

1.1. Company information

The company that I will do my bachelor thesis for is ProRail B.V. (ProRail). ProRail is responsible for the rail network in The Netherlands (ProRail, 2019). In short this means the maintenance, security, management and the instalment of the railroads.

The department I work on, mostly deals with the maintenance part of ProRail. The department is called Operationeel Besturingscentrum Infra (OBI), loosely translated into Operational Control Centre Infrastructure. OBI is a part of the control centre of ProRail as can be seen in *Appendix A: Organization chart.* This appendix gives an overview of where OBI belongs to in the company.

On the department of OBI, the energy supply of the conventional network, the HSL (high-speed track) and Betuweroute (cargo rail track) and the main feed of switches, traffic lights, catenary etc. are monitored and operated. Next to this, the OBI keeps an eye on different tunnel technical installations in The Netherlands and they work on all infra-related failures and coordinate the completion of them. These failures differ from a broken switch to a broken lamp in a traffic light (Infrabeschikbaarheid ProRail, 2019).

OBI is divided into 4 work lines, also called stations: North 1500V, South 1500V, South 25kV and day shift. North 1500V deals with the northern part of the Netherlands, South 1500V with the conventional network in the southern part of the Netherlands and South 25kV deals with the Betuweroute and the HSL. The day shift is an extra shift that does the preparation work for the scheduled maintenance jobs. They prepare so called switch sequences that have to be executed before and after the maintenance.

1.2 Action problem

ProRail came to me with a problem regarding the employee roster of the OBI. They thought that it was possible to schedule the employees more efficiently, such that less employees were needed per week. With this starting problem in mind, I started looking for all problems at the OBI. All problems were put into a problem cluster, to get the causal relations (*see Appendix B:*). In the next paragraph, the detailed problems will be explained as can be seen in the cluster.

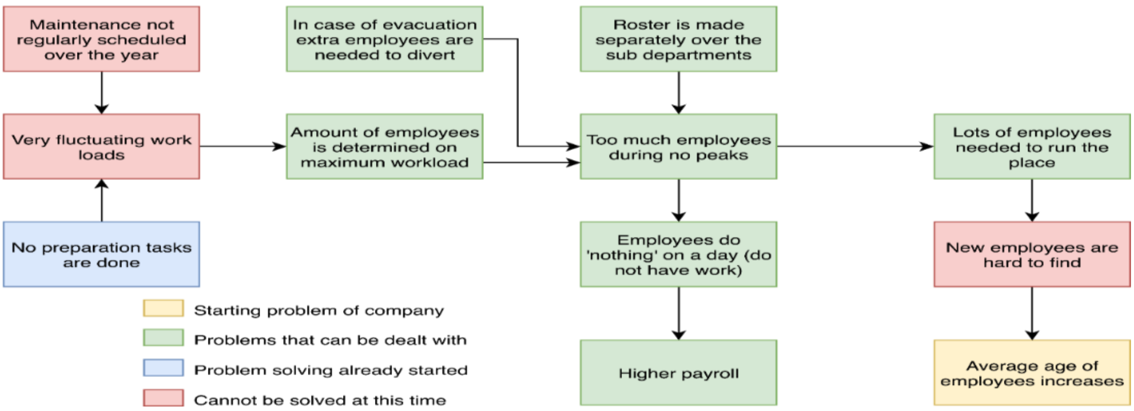

Figure 1: Problem cluster

1.3 Problem description

OBI works 24/7 in eight hour shifts. On every work line there is at least one person at the desk to handle possible failures etc. This rule counts for the North 1500V, South 1500V and South 25kV stations. If every employee works 5 shifts a week and a day consists of 3 shifts, this means that 12.6 full-time employees (FTE's) have to be at the stations. Currently 2 employees are scheduled on the North stations. This results in ever higher amount of FTE's needed every week, namely 16.8 FTE's. The average age of the employees is high and increasing, therefore ProRail needs new employees. Acquiring employees is very tough since there are little potential employees on the labour-market.

Currently the schedule is based on the maximum workload that could occur. This workload is very fluctuating due to irregular maintenance schedules and preparation tasks which are partly done. This results in a schedule which is based on the maximum amount of maintenance and little preparation tasks done. A result of this is that too many employees could be scheduled in for a shift. With the increasing age of the employees and the toughness of acquiring employees, the chance of employee shortage is increasing.

1.4 Restrictions

In this thesis, only one problem can be handled. As explained by Heerkens and Winden (2012), the first problem in the cluster is the most beneficial to be handled, as that problem is the cause of multiple other problems. However, as can be seen in the problem cluster (*Appendix B: Problem cluster),* not all problems can be fixed. To improve the readability of the next chapter, the problems which are at the beginning of the cluster will be summed up.

1. Maintenance not regularly scheduled over the year
2. Little to no preparation tasks are done
3. Fluctuating workloads
4. Amount of employees is needed is scheduled on possible maximum workload
5. In case of evacuation extra employees are needed to divert

The irregular maintenance (1) cannot be changed. This is due to legal contracts ProRail has with the contractors. It is stated in the contracts that contractors can plan their maintenance at any time they want. As ProRail cannot influence this right now, this problem cannot be fixed and thus can never be the core problem (Heerkens & Winden, 2012).

The low amount of preparation tasks (2) which are done right now, is researched now. So for the company it is of no need to let me research that as well. The fluctuating workloads (3) is a result of the first 2 problems and cannot be handled if the those problems cannot be solved. This results in two problems which are left to choose as core problems. First of all, **the amount of employees which is determined on the maximum workload possible (4).** Secondly **the case of emergency evacuation for which extra employees are needed (5).**

This thesis will focus on the scheduling of employees on the maximum workload. This problem is every day, whereas evacuations rarely happen. Therefore solving the scheduling is more beneficial for the company.

1.5 Research methodology

There are several methodologies to solve business problems. At the University of Twente, the most used methodology is the Managerial Problem Solving Method (MPSM) by Heerkens and Winden (2012). This methodology consist of 7 phases, namely:

1. Defining the problem
2. Formulating the approach
3. Analysing the problem
4. Formulation (alternative) solutions
5. Choosing a solution
6. Implementing the solution
7. Evaluating the solution

The MPSM will be used as a framework. Whenever needed, adjustments will be made to the method. These adjustments are easy to implement, so the method can be used in varying studies. When using the MPSM, the problem will be expressed in terms of variables, which will be measured.

Due to the measurability of the variables, the improvement of the solutions can be calculated. By weighing the importance of the variables with the improvement, a solid solution can be made.

The first step of this methodology is already done in the first part of this chapter. Therefore the research question can be formulated. A research question consists of the reality and a norm (Heerkens & Winden, 2012). Between the reality and the norm is a difference, thus there is a problem. The reality in this case is that there are too much employees scheduled per shift. The norm is as little employees as possible. This is easy to measure, by summing up the total number of employees per week. This results in the following research question:

Is there a way of scheduling the employees more efficiently, such that less employees are needed to perform the same amount of work?

1.6 Approach of the problem

This section will describe the approach to come to a solution for the research question. To be able to come up with a solution, knowledge has to be gathered. To be sure which knowledge to find, sub-questions are made, these can be found in chapter 1.6.1. The approach to solve the sub-questions is given as well.

Sub-questions

Below the sub-questions are stated. Every sub-questions is divided into multiple smaller questions. The questions are made in such a way, that together with the knowledge I already have myself, I am able to come up with a solution to the research question. After the sub-questions a short explanation is given on how I am going to get answers to the questions.

1. **What is the distribution of the calls, maintenance jobs, switch sequences and failures that could occur? What if no distribution can be found?**
    - **a.** How many calls do the employees get per shift?
    - **b.** What is the duration of a call?
    - **c.** How many maintenance jobs are scheduled?
    - **d.** Do all maintenance jobs need a switch sequence?
    - **e.** What is the duration of the making of a switch sequence?
    - **f.** What is the durations of checking the switch sequence?
    - **g.** What is the distribution of the amount of failures?
    - **h.** How long does it take to handle the failure?

In question 1 distributions will be found. To be able to look for potential distributions, first of all the datasets have to be gathered. As almost all data is saved in the system, historic data can be used as datasets. These datasets will be used to answer the questions a, b, c, e, and f. For question d the employees can be asked. They will know whether sequences are needed or not.

Question g and h will be harder to solve. These data cannot be found in systems and it is very hard to observe, as the chance of occurring is so low. However, because the impact of them is large, they should be added in the model. The best way of answering the questions is by asking multiple employees to assess.

In the end this question will give a method to look for distributions and all the distributions for the datasets. These can be used as input for the mathematical model.

2. **Literature study: What mathematical models can be used to calculate the number of employees needed? And how to optimize according to this model? Or what algorithms or heuristics are used to solve employee scheduling problem?**
    - **a.** Which mathematical model is suggested in the literature to optimize schedules with uncertainty in the input?
    - **b.** What are the preconditions, assumptions and restriction of those models?
    - **c.** What are the preferences, restrictions and limitations from ProRail?
    - **d.** Which models can be used to optimize the number of employees at OBI with regard to the results of the questions above?
    - **e.** Is there a way to get deterministic numbers out of the data?

This question will be the basis of this thesis. Out of this question, the solving method will be determined and set up. To solve this question, existing literature will be analysed to find possible mathematical models to use. Furthermore, other ways, like algorithms or heuristics will be looked at. For every model, the preconditions, assumptions and restrictions will be written down.

3. **How to make sure that all restrictions are met in the model? (As there is a problem that not everybody may work night shifts or cannot work on all work lines)**
    - **a.** What are the restrictions?
    - **b.** Is it necessary to put all restrictions in the model? Or can some be left out? (simplifications)
    - **c.** How to adjust the input data such that all chosen restrictions are met?
    - **d.** Will the addition of restrictions affect the validity of the model?

By talking to the employees, observing myself and talk with the management, the restriction will become clear to me. Whether I have to put them in the model will be discussed during a meeting with the management. Furthermore, other thesis can be used to look how they deal with restrictions.

4. **What KPI's will be chosen to determine the performance of the system?**
    - **a.** What KPI's can be measured?
    - **b.** Which of those KPI's can work in this specific case?
    - **c.** Which of the KPI's are very important and which are less important?

For this question my own perspective as well as the perspective of the managers will be taken into account. Existing literature on the same topic can be analysed as well. In that way no KPI will be missed. Together with the management the most important can be chosen.

1.7 Research objective

The objective of this research is to look into possibilities to decrease the number of employees per shift. To come to this, some steps have to be made. First of all, the workload per work line has to be determined, divided into the shifts. Secondly, all restrictions regarding the scheduling of employees, the workload or the occupation have to be stated. By combining the workload with the restrictions, the minimal number of employees needed can be calculated by making use of a mathematical model. By putting the output of the mathematical model into a simulation model, the solution can be verified on whether it is achievable for the company.

1.8 Scope

This research will focus on improving the number of employees per shift. Due to time constraints, the input will be divided in four types: incoming calls, making of switch sequences, checking of switch sequences and the failure handling. Further division of the kind of calls or failures will not be made, as the difference in duration is minimal and this will increase the difficulty of the model. This thesis will not focus on making a schedule, but only on the number of employees. Next to this the current situation (see sub-question 1) will be analysed as well as some other scenarios the company asked me to analyse (stated in research objective), if there is enough time left. The main priority lies on the current scenario.

1.9 Deliverables

The deliverables of this research are summed up below. The main report should be readable without opening the models, so all models which are built are explained in there, together with the solutions.

- A model that gives insights in the amount of employees to schedule per shift (mathematical model). Reusable in the sense that I can use it on different scenario's by only changing the input.
- Simulation model that can be used to verify the outcomes of the mathematical model.
- Detailed report regarding the steps followed to come to the conclusions. Gives insight in the qualitative and quantitative analysis done.
- Conclusion and advice to the company

1.10 Limitations

This research has to deal with some limitations. First of all, the 10 weeks' time constraint. This research has to be finished in 10 weeks. This means that not everything can be done as detailed as it could have been done. In the following chapters, this constraint will be mentioned multiple times, since it affects the research a bit. Secondly, some limitations are given by the company. The company asked for a model that at least could show the flow of the jobs through the system with different numbers of employees, therefore a simulation model is made. Regarding the general limitation there are no more, however, in the next chapters other limitations will be made which focus on specific parts of the report, therefore it is stated in the chapters.

Chapter 2: Literature review

In this chapter, the sub-questions will be answered by searching in existing literature. This chapter will give insights in the several workforce scheduling models to use (2.1), mathematical models (2.2), which key performance indicators to use (2.3) and what kind of simulation models there are and how to make sure they are valid (2.4).

2.1 Workforce scheduling

In this research, workforce scheduling will be done. Workforce scheduling deals with the planning of personnel to shifts in order to cover the varying demand. The amount work is often irregular and thus the number of employees needed differs over time. Typically, various constraints makes the scheduling complex (Pinedo, 2009). In literature there are 3 basic methods, namely:

- Day-off scheduling heuristic
- Shift scheduling
- Crew scheduling

Day-off scheduling

First the day-off scheduling method will be considered. This model uses the minimum capacity needed at the company as well as regulations regarding workweeks. This means that this schedule will roster enough employees considering the amount of days an employee can work at maximum per week, the number of weekends they can work per month and the amount of rest they need between shifts.
So the general approach in this model is to first determine the restrictions the model should meet. Next these restrictions will be stated in mathematical constraints, which can later be put into a model. After putting all of this in a model, this model can be solved. However, as these models take often too long to solve, the solving algorithm is only executed for a certain duration. After that, the heuristic is used to come up with a solution as close to the optimum or the optimum.

The overall objective of this method is to minimize the number of employees scheduled per week (Pinedo, 2009).

Shift scheduling

This scheduling method has the objective to minimize the total costs (Pinedo, 2009). Assigning an employee to a shift costs money. The amount of money it costs differs per shift. Every employee is assigned to only one shift. Think of different shifts like day and night shifts. Night shifts often cost more as the employees get a bonus for working at night.
Every hour per day a different amount of staff is needed, however, there are strict shifts. The objective is to assign the employees to the different shifts to make sure that all the demand is met and the costs are as low as possible. In the book Planning and Scheduling in Manufacturing and Services (Pinedo, 2009), a good example is given of a retail store. The employees can work five different shifts per day. By combining the different shifts to the demand per hour, an optimal solution can be found.

Crew scheduling

This part of workforce scheduling is the scheduling of employees to task. Every task is assigned to a trip and an employee is assigned to the trip. These scheduling problems often occur at transportation companies. The goal is to assign each task to a trip in such a way that the total distance travelled is as low as possible and thus the costs are low.

Conclusion

All of the methods make use of a similar build up to come to the solutions. This build up can be explained step-by-step as follows:

1. First the case has to be made in detail, so the constraints have to be determined, the number of employees, the workload etc.
2. The findings in step 1 have to be put into a model, mostly a Linear program (LP) model. So everything has to be put into variables. After that has been done, an optimization formula has to be derived, which will give a solution to the problem.
3. After the model is made, it has to be solved. As solving optimally takes a lot of time, the model is solved until one solution has been determined. This solution is the 'best solution' for now.
4. With the solution of step 4 as initial solution, algorithms and heuristics, like the Day-Off heuristic will be used, as well as regression on all the possible neighbour solutions. These neighbour solutions can be determined by local search.
5. After finishing step 4, all outcomes can be analysed and the best outcome will be chosen as the schedule.

Step 1 is the analysis of the current scenario at the department. This current scenario analysis can be found in *Chapter 3: Current situation*.

Chapter 4: Data analysis and *Chapter 5: Solution design* are used to determine the data parameters and to explain the mathematical model. This is all needed for step 2. Step 3 and 4 will be done separate to this model, but all steps and results will be explained in *Chapter 6: Results* and *Chapter 7: Conclusions and advice.*

2.2 Mathematical model

This thesis will make use of a mathematical model. Mathematical models describe beliefs about how the world functions into the language of mathematics (Marion & Lawson, 2008). The objective of the mathematical model will be to test the effect of changes in a system. The changes in this case are the number of employees at the department.

Within mathematical models, there are multiple broad classifications. These classifications tell immediately some of the essentials of the structure of a model (Marion & Lawson, 2008). The most essential classification for this thesis is the difference between deterministic and stochastic models.

Deterministic models are models in which the outcome is always the same, no matter how often the model is solved. There is no random variation within the model, and through relationships among variables, which are known, outcomes can be precisely determined (BusinessDictionary, 2019).

Stochastic models are models in which the outcome relies on some kind of randomness. This means that probabilities are used to come to a solution. Every time the model runs, the output will likely differ (Stephanie, 2016).

Stochastic models are in most cases harder to solve, but give more precise solutions, especially in small samples (Stephanie, 2016). A distinction has to be made between these two terms.

Stochasticity can be described in many ways. In stochastic models, the full stochasticity will be used. Thus a distribution will be found and later on used in the model. However, there are other ways of getting as much stochasticity in a model, without having to use a stochastic model. One of this ways is to make use of smaller bins. By using smaller bins, the average will be determined more often. This means that fluctuations have much more influence on the model than by determining the average over the whole dataset. By determining the means, a deterministic model can still be used.

Furthermore, in the case of employee scheduling, the difference between stochastic and deterministic solutions will be less. This is because the difference between scheduling 2 or 3 employees is in most cases larger than the difference in solution of the different models (with decimal numbers).

Because of the small difference in solution and the extra work and complexity it is to make a stochastic model, it is chosen to make use of a deterministic model.

Types of models

There are several types of mathematical models, like dynamic programming models, discrete or stochastic models. The most common ones are the linear programming models (LP models). Since this research has to calculate the number of employees, an Integer Linear programming model is the only variant of the LP models that will work. In this thesis indeed an ILP will be used to determine the number of employees (see *Chapter 5: Solution design*).

For the solving of the model several software programs could be used. Think of Excel, R, Matlab or Lingo. For this research Excel will be used as solver. The reason to choose Excel above the other programs is because ProRail uses Excel. For them it could be useful that the model is reusable for a small part. Since Excel is capable enough to solve it, no problems will occur.

2.3 Key Performance Indicators

The department of OBI is a control centre. When something goes wrong, they have to react very quickly. Especially within the rail sector the stakes are very large. Every failure affects a lot of travellers. Therefore it is for utmost importance that tasks are handled very quickly. A way to measure the handling speed of the department is by using the waiting time of a job in the system as a key performance indicator (KPI). This KPI is determined together with the management team.

In literature the most common KPI for such kind of problems is the number of employees. This is straight forwarded, since this is also the objective of the study. Therefore this KPI is added to the model.
Next to this it is also important that an employee is not utilized for 100%. That would mean that no extra work or delays could be permitted. This KPI was a wish from the department, since they do not want to be overloaded with work. The KPI utilisation is therefore added as the last KPI.

- Waiting time of a job in the system
- Utilisation or workload per employee
- Number of employees rostered

Waiting time of a job

The meaning of this KPI is literally said in the title. This KPI is the time between a job coming in and the job getting started. This time will be averaged over all the jobs to come up with a number. The lower the number, the less waiting time per job there is on average.

Utilisation or workload per employee

This KPI looks at how much work an employee has to do in its shift. The higher the utilisation (time an employee is busy), the less employees are needed. However, if the utilisation is very high, a lot of pressure is on the employee. Too much pressure can lead to illness, therefore a maximum has to be set on the utilization.

Number of employees rostered

The objective of this thesis is to schedule as less employees as possible. Therefore, the total amount of employees rostered is very important to keep track off. This is the KPI that can be changed during this thesis.

The first KPI is unable to measure in a deterministic model. In a deterministic model, only the average amount of work that can be done in an hour is taken into account. The time a job waits is not taken into consideration. Therefore next to the mathematical model, a simulation model has to be made. This simulation model will be able to calculate per job the waiting time and is able to calculate the average over all jobs.

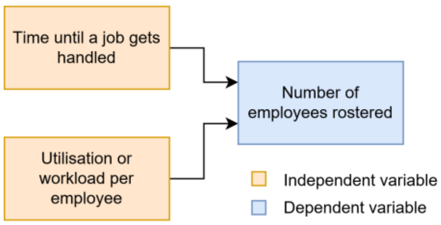

Figure 2: KPI model

2.4 Simulation model

"Simulation studies are computer experiments that involve creating data by pseudorandom sampling. The key strength of simulation studies is the ability to understand the behaviour of statistical methods because some ‘truth’ (usually some parameters of interest) is known from the process of generating the data.” (Morris, White, & Crowther, 2017). This quote out of the work of Morris, White and Crowther describes very well what a simulation study is. In short a simulation model is a recreation of the reality to be able to value different chances in comparison to the reality.

The first step of making a simulation model is creating a conceptual model. Figure 4 gives a schematic overview of what the steps are during conceptual modelling. The conceptual model consists of the input for the model, activity diagrams, assumptions, simplifications, restrictions and output of the model. In fact the conceptual model is the complete description of what to use and what to calculate during simulation.

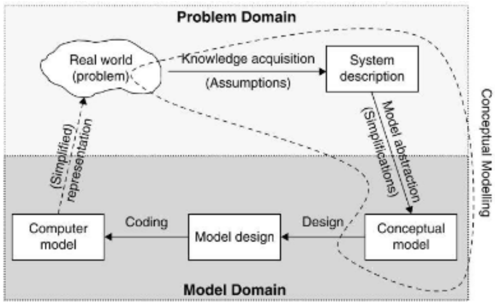

Figure 3: Build up simulation model (Robinson, 2014)

The step after the conceptual model is the model design. This consist mostly of collecting data and finding distributions to fit this data. After finishing this step, all the flows of the jobs through the system are known together with the distribution for the duration and the number of jobs. This is all the data needed for the model and thus the coding can be done.

The coding can be done in many different software programs. Think of Simul8, AnyLogic, SimScale or Tecnomatix Plant Simulation (TPS). For this thesis TPS will be used, as this software is known to me.

Nature of simulation models and simulation output

Simulation models can be terminating or non-terminating simulations (Robinson, 2014). Terminating simulations have a natural endpoint. This can be the closing time of a shop or the end of the time-period of the investigation or the completion of a schedule. (Robinson, 2014). Non-terminating simulations will never stop. “There is no reason for the simulation to stop other than the user interrupting the run.” (Robinson, 2014). The length of the simulation has to be determined by the user for these kind of simulations.

The department of OBI is a 24/7 active control centre. This means that the simulation should be a non-terminating simulation.

Non-terminating simulations often reach a steady-state (Robinson, 2014). Steady-state means that the output of the KPI’s will vary according to some sort of a distribution, called the steady-state distribution. This steady-state can be two forms. It can be a steady state point/line, meaning that after reaching this steady-state the output will vary according to the point/line. Another option is the steady-state cycle. This means that there is a certain pattern that keeps repeating itself.

Warm-up length

For the validity of the output it is important to know the warm-up length. This can be determined by making use of the Marginal Standard Error Rule (MSER) or a graphical method (same as replications graphical method). “MSER has the aim to minimise the width of the confidence interval about the mean of the simulation output data.” (Robinson, 2014).

$$MSER(d) = \frac{1}{(m-d)\cdot 2}\sum_{i=d+1}^{m}\left(Y_i - \bar{Y}(m,d)\right)^2$$

d = the proposed warm-up period
m = the number of observations in the time-series of output data
$\bar{Y}(m,d)$ = the mean of the observations from Y_{d+1} to Y_m

For every *d*, the MSER will be calculated. This means that for every *d,* the difference between the current mean and the mean of all the data coming up later is determined. The *d* which has the lowest MSER will be the warm-up period. Reason for this is that the difference of the current KPI value and the value in the future of the simulation is the lowest.
There is one constraint to the determination of the warm-up period. If the warm-up period is more than half of the run length used to determine the MSER, this value is not accountable and the MSER has to be determined out of a larger data sample.

Replications

Not only the warm-up period is important to know, also the length of the run and the amount of replications of the run are important. Running the simulation too short can cause that the data is too little to make proper analysis out of. However, running the simulation very long increases the run-time. The same counts for the number of replications. The more replications, the more data there is, but the runtime increases as well. Therefore, for both terms, the distinction has to be made between runtime and the amount of data.

"A replication is a run of a simulation model that uses specific streams of random numbers." (Robinson, 2014). By changing the streams of random numbers, more randomness is inserted in the model, resulting in a more accurate output.

Robinson (2014) describes three approaches to come up with the amount of replications, namely: a rule of thumb, a graphical method and a confidence interval method.

A rule of thumb

Robinson (2014) states that according to Law and McComas (1990) at least three to five replications are needed. This method does not look into the output of a model, but is just an advice that could be followed.

Graphical method

The graphical method uses the cumulative mean of multiple replications to determine the amount of replications needed. The first step of this method is to make a lot of replications of a simulation model (around 20). For every of these replications, the runtime is the same and the warm-up period is used. Every replication gives a KPI value. A graph can be made with all cumulative means, one mean for every amount of replications. The point at which the graph is nearly flat, is the amount of replications needed.

Confidence interval method

This method is an addition to the graphical method. Instead of plotting the cumulative means, a confidence interval is made. The goal of this method is to achieve a deviation percentage lower than 5%. In the book of Robinson (2014), this method is explained in detail. This method uses formulas to determine the actual deviation for every number of replications.

Because of the 10 week time constraint and because of the little difference between the confidence interval method and the graphical method, it is chosen to use the graphical method for both the replications as well as the warm-up length. This method is a lot quicker to determine and almost as accurate as the confidence interval method/MSER. Because of the decision to use the graphical method, the confidence interval method is only shortly described, as it is of no usage for this research.

Run length

After determining the number of replications, the run-length should be determined. There is almost no information on how to calculate the run-length. However, according to Banks et al. (2009), the run-length should at least be 10 times the warm-up period. Therefore this will be used in the thesis.

Chapter 3: Current situation

This chapter describes in 3.1 the different stations and the current number of employees per station. Chapter 3.2 will explain what kind of work the employees have to do per day. Chapter 3.3 will describe what adjustments are made to the study, regarding the current situation and chapter 3.4 gives an insight on how to solve this problem.

3.1 Stations

The department of OBI is divided into 4 stations (North 1500V, South 1500V, South 25KV and Day shift). All stations together have the objective to maintain and manage the electricity network around rail tracks. All stations have a different tasks to reach this objective. It is useful to understand the difference and therefore a clarification per station is given below.

Figure 4: Station in the Netherlands

North 1500V

North 1500V is the station that keeps track of the electricity on the rail tracks in the northern part of the Netherlands. The northern part is from Leiden to Utrecht to Winterswijk and everything above. In *Figure 4*: Station in the Netherlands this is the black line and everything north of that line. 1500V means that the electricity on the catenary is made out of 1500 Volt. This is the standard in the Netherlands. So the conventional rail tracks in the Netherlands are 1500V.

South 1500V

South 1500V is the station that keeps track of the other part of the Netherlands. So all the conventional rail tracks in the southern part of the Netherlands. Below the black line.

South 25kV

South 25kV is the station that manages all the rail tracks in the Netherlands that make use of 25000 Volt as input for the under stations. Nowadays all the rail tracks using this system are in the southern part of the Netherlands. The tracks using this system are the “Betuweroute” and the High-speed track (HSL). The technique used for 25kV is very different from the 1500V technique, therefore, these stations and employees cannot be combined with the 1500V stations.

Day shift

The day shift is separated into North and South, but for the easiness, both are combined into the day shift work station, as the work they do is the same. During the day shift, the sequences are made so that when maintenance starts, the electricity on the specific parts of the rail track can be switched off easily. This shift is temporary. In the future, the sequences will be made automatically, such that this shift is unnecessary then.

The current number of employees per shift at a station can be seen in ***Error! Reference source not found.***.

Table 1 Number of employees per shift

Station	# of employees
North 1500V	2
South 1500V	1
South 25kV	1
Day shift	2

3.2 Work activities

OBI has in broad lines two types of activities. Activities that occur ad hoc and planned activities. This division will be used in this chapter to clarify all the activities.

Planned work

There is only one type of planned work, namely planned maintenance. The work activities related to this planned maintenance are the making, checking and performing of switch sequences.

A switch sequence is an order in which parts of the electricity network will be switched off or on. With parts, the electricity switches or overhead wires are meant. The normal procedure with switch sequences is as follows: one employee makes the sequence in the way he thinks is necessary. This making is done in the day shift. The employee performing this sequence at night, checks it in that specific night shift. If he agrees, he performs it the way it is in the system. Otherwise he performs it himself in another way. *Figure 5*: Flowchart of the system shows the sequence activities in a flowchart. Out of the 3 different flows in figure 6, only the sequence flow belongs to planned work.

The variability within this input is in the number of sequences per night, as well as, the complexity and the size of the sequence itself. This complexity cannot be predicted. Complexity of sequences is dependent on the location of maintenance. The amount of electricity stations, the number of switches and the number of rail tracks can improve the complexity very quickly. Combining the location with the length of the rail track on which maintenance is performed, makes it almost impossible to predetermine the complexity level and thus the time the checking costs.

Ad hoc work

The ad hoc work can be divided into three big activities, handling incoming calls, solve major failures, and manual switch handlings plus some other smaller activities. First of all the big activities will be explained.

Employees get a lot of calls per shift. This can be calls to ask permission to go into an electricity substation or calls to ask for a switch to be turned on or off. The division within the type of calls can be made very large. There are over 10 different types of calls. Due to the fact that most call types take about the same amount of time, these all can be combined into one input type, namely the calls. This saves a lot of time, because the data analysis is shorter.

With major failures, broken catenary or anything that causes electricity systems to not function properly is meant. The major failures are very hard to predict. There are so many things that can break down in the whole catenary. ProRail has started a lot of projects to improve the predictability of failures, such that preventive maintenance can be used more effectively. However, currently this is not good enough yet.

In both cases there is variability. The calls have a chance of occurrence and a chance of having a certain duration. The same counts for the failures. The failures have a very high variability as the duration and the amount of failures differs a lot per week.

The third big activity is the manual switch handlings. These are the switches they perform manually and mostly to check whether switches for example work during or after maintenance. The problem with this data is that it is in the logbook, but very hard to get it out of it. Only the professionals can see the different manual switch activities, however, there was no possibility for them to get it out of the logbook. What can be said is that most of these activities cause some more phone traffic. The more phone traffic, the more chance for manual switch activities and the more time there is used to manually switch.

The other smaller activities are taking over the work of the "Coördinator Herstel Infra" (CHI) during the night shifts and weekend evening shifts answering questions of the employees of the Backoffice. The CHI works via calls which are connected to mobile phones. The problem with the connection towards the mobile phone, means that it is not logged in the voice logbook. This results in no data. Furthermore, the number of calls for the CHI cannot be told by the employees since it differs a lot. Therefore, these extra work will be added in the conclusion and advice, when the results are being analysed.

The flowchart of all the activities can be seen in figure 6. There are 3 different types of input, namely calls, failures and sequences. Whenever a job comes in, no matter what type, it goes into the sorter (queue of jobs) and whenever an employee is free, it handles the job and the job leaves the system.

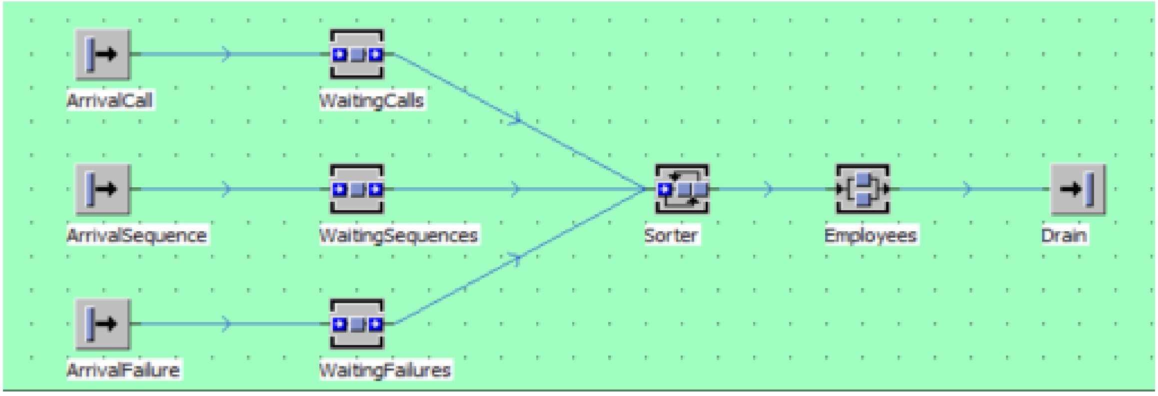

Figure 5: Flowchart of the system

One difference that cannot be seen in this flowchart, is that sequences follow this route 3 times. Due to the fact that sequences have to be checked once and executed twice, in total they follow this flowchart 3 times.

3.3 Adjustments to the model

Due to the current scenario, some adjustments will be made to the model and the scope of the model. First of all the working stations. The stations that will be analysed in this research will decrease to 2, namely North 1500V and South 1500V. The reason to drop out South 25kV is because there is only 1 employee at the desk and there is no room to decrease this number as there has to be always one employee present. Logically, this would also count for South 1500V, as there is also 1 person per shift. Why will South 1500V be analysed? The decision is made to check whether the combination of North 1500V and South 1500V into Netherlands 1500V is possible and beneficial. Therefore, South 1500V has to be analysed.

Another reason to leave out the stations is because of the time-constraint of 10 weeks. By decreasing the number of stations to analyse, the constraint can be met.

The day shift is left out, because this shift is temporary. The automation of the tasks is started. In short time, the computer will take over the human work. So in discussion with the management team and my supervisor of ProRail, it is decided to leave out this station and focus fully on the other two stations.

The tasks are all checked with data on how often they occur. This resulted in the decision to leave out the work of the CHI and the questions by the Backoffice employees. The data showed that the amount of work per shift on these tasks is about 2 minutes on 8 hours, so negligible.

3.4 Flowcharts

To make all the adjustments clear and more understandable, flowcharts have been made. These flowcharts in figure 7 show all the input data, the models and the output that will be generated.

This flowchart will be followed three times. One time for the North 1500V station separated. One for the South 1500V station. The last one will be the Netherlands 1500V scenario, which is the combination of North 1500V and South 1500V. There is no need to do North and South into one model but keep them separate, since this will not result in other solutions as they are completely separated.

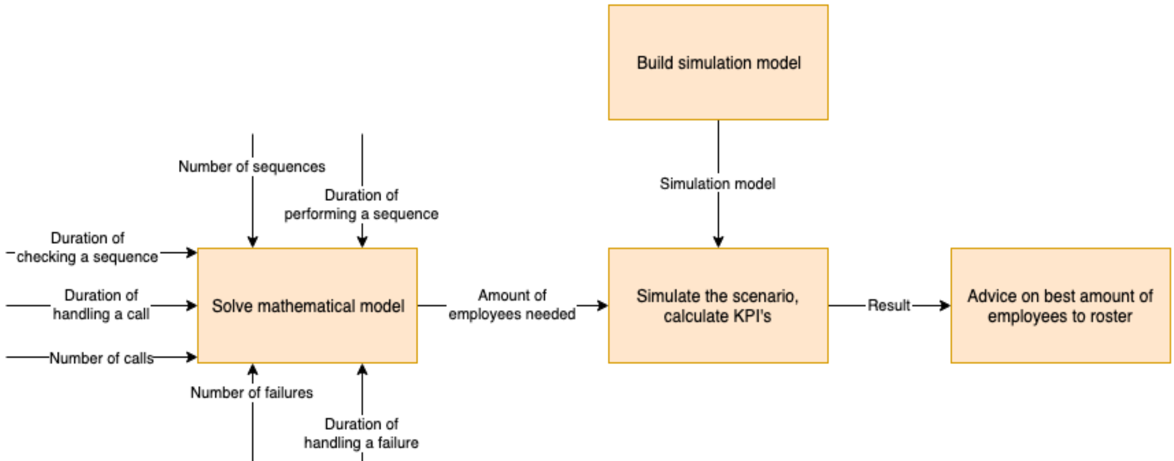

Figure 6: Flowchart model building

The flowchart of figure 7 is divided into three parts. First of all the data analysis has to be done. This data analysis will determine all the input distributions with their means and variations. This data analysis is all explained in

Chapter 4: Data analysis. The second part is the building and explanation of the mathematical model and the simulation model. This is explained in *Chapter 5: Solution design*. The third and last part is the evaluation and the conclusion of the models. This is done in *Chapter 6: Results.* In chapter 6 all the KPI's of the current system can be found.

Chapter 4: Data analysis

This chapter describes in section 4.1 the data which is needed, in section 4.2 how to gather the data, chapter 4.3 shows the way how to find the distributions by using R. Chapter 4.4 states extra assumptions and explanation on some distributions. The last chapter 4.5 gives extra information for the result of the data analysis. All the important parameters are stated there.

4.1 What data is needed?

To come up with a solution, all data about the number of activities and the duration of the activities should be known. There are a lot of different activities at the department. Not all activities can be analysed separately due to time constraints. Therefore, assumptions and simplifications have been made to the activities. Activities have been grouped according to their duration. Activities with about the same duration have been put together, reducing the amount of datasets.

The activities have been grouped into the following groups:

- Handling the incoming calls;
- Executing the switch sequences;
- Checking the switch sequences;
- Handling the disruptions in the electricity network.

So for every group, the number of occurrences has to be known as well as the duration of the activity.

4.2 Available data

The department makes us of a software system to monitor and change the electricity network. This software makes logbooks and also saves them. This means that every handling of the employee is logged.

These logbooks have been gathered for a period of 0.5 years. Personnel of the department said that this period should be representative for the full year, with exceptions of the holidays. Out of the logbooks, the number of sequences as well as the duration of performing them can be determined.

The checking of the switch sequences is not stated in any document or software programme. Therefore, the only way to gather data is by observing the personnel. However, these observations have to be done at night and due to public transport issues at night, there was not much time for observations. Therefore informal conversations have been done with some employees, resulting in a rough estimate of the duration of checking switch sequences.

For disruptions, the same counts as for the checking of sequences. There is no data available stating the amount of failures. Therefore there are two other options. Observation can be used, however, it is unclear when disruptions happen. This could mean that during observations nothing may happen, which may lead lost of time. Therefore, a rough estimate is made together with the employees, just as with the checking of switch sequences.

The handling of incoming calls is determined by historical data. Every call is voice logged into a software programme. By extracting the time of the calls and their duration, the distributions can be determined.

4.3 Determination of distributions

Within the domain of statistics, there are many ways to determine distributions. For this thesis, only one of these methods is used to determine whether a distribution can be found for the data.

To determine the distributions, a statistical programme or software will be used to easily calculate and plot the data and distributions. This software will be R. At the website of R (R Core Team, 2019), the programme is described as: “R is a system for statistical computation and graphics. It provides, among other things, a programming language, high level graphics, interfaces to other languages and debugging facilities.”

In R, the package ‘fitdistrplus’ will be used. This package is made specifically to determine distributions of datasets (R Core Team, 2019).

The method to determine the distributions, will be the same as described in the paper of M.L. Delginette-Muller et al. (2009). This method consist of the following steps:

1. Graphical display of the dataset
2. Characterization of the dataset
3. Fitting of a distribution to the dataset
4. Simulation of the uncertainty of the distribution to the dataset

The above steps is the general way of finding a distribution. However, it is not clear yet, what to do at each step. Therefore, per step, a detailed explanation will be given on how every step will be performed. Within the explanation, the small lines of code used in R will be stated.

Step 1: Graphical display of the dataset

The graphical display of the dataset will be given by plotting a histogram as well as plotting the cumulative distribution of the data. These graphs will give an insight on how the data is divided. This can give already some ideas of which distribution might fit to the data.

```
> data <- c(“dataset”)
> plotdist(data)
```

Step 2: Characterization of the dataset

In this step, the descriptive statistics of the dataset will be determined. This includes the: minimum, maximum, median, mean, standard deviation, skewness and the kurtosis. Next to this, the Cullen and Grey graph will be plotted. This graphs shows, by plotting the square of the skewness and the kurtosis, potential distributions that lie close to the dataset.

```
> data <- c(“dataset”)
> descdist(data)
```

Step 3: Fitting of a distribution to the dataset

This step is closely related with step 2. In step 2, the Cullen and Grey graph is plotted. As said, this gives potential distributions. In this step, in order of closest related to less related, the distributions will be fitted to the data. This means that the data will be plotted together with the distribution in one graph. Next to this graph, the QQ-plot, PP-plot and cumulative distribution functions will be shown. All these graphs have the objective to show whether the distribution is representative for the dataset. Furthermore the rate and the shape will be determined.

If a distribution is closely related on sight, a goodness-to-fit test will be used. For continuous distributions, the Kolmogorov-Smirnov test will be used. For discrete distributions, the chi-square test can be used. The Kolmogorov-Smirnov test is as follows in R:

```
> data <- c("dataset")
> ks.test(data, "distribution", "parameters")
```

Next to the test, also the specific parameters of the distribution will be determined. This is done with the following code:

```
> data <- c("dataset")
> fig <- fitdist(data, "distribution")
> plot(fig)
> summary(fig)
```

During the Kolmogorov-Smirnov test a p-value will be calculated. If this p-value is larger than 0.05, the H_0 hypothesis cannot be rejected and thus the chosen distribution represents the dataset in a good enough way.

Step 4: Simulation of the uncertainty of the distribution to the dataset

The last step is about bootstrapping the dataset. This gives per data point the confidence interval it can lie in and also plots the bootstrapped values. The summary gives the median as well as the 95% confidence interval of the dataset.

```
> data <- c("dataset")
> fig <- fitdist(data, "distribution")
> big <- bootdist(fig)
> plot(big)
> summary(big)
```

Conclusion

To conclude, step 1 to 3 determine the distribution which will represent the dataset. If the p-value is larger than 0.05, the distribution can be used as representation. Furthermore, by doing step 4, a confidence interval can be made of the parameters. This gives, together with the graphs, a good overview of how the variability is, when random generating values multiple times. This can later be used to adjust the parameters to more extreme values (more work, so less risk of under capacity).

4.4 Determination of distribution

As explained in *2.4 Simulation model*

*"*Simulation studies are computer experiments that involve creating data by pseudorandom sampling. The key strength of simulation studies is the ability to understand the behaviour of statistical methods because some 'truth' (usually some parameters of interest) is known from the process of generating the data." . This quote out of the work of Morris, White and Crowther describes very well what a simulation study is. In short a simulation model is a recreation of the reality to be able to value different chances in comparison to the reality.

The first step of making a simulation model is creating a conceptual model. Figure 4 gives a schematic overview of what the steps are during conceptual modelling. The conceptual model consists of the input for the model, activity diagrams, assumptions, simplifications, restrictions and output of the model. In fact the conceptual model is the complete description of what to use and what to calculate during simulation.

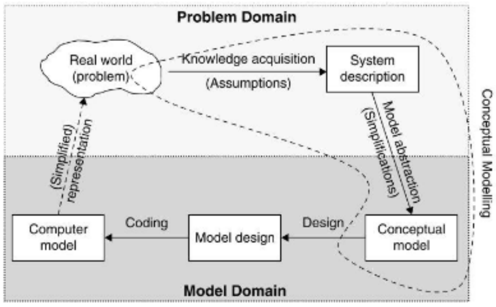

Figure 3: Build up simulation model (Robinson, 2014)

The step after the conceptual model is the model design. This consist mostly of collecting data and finding distributions to fit this data. After finishing this step, all the flows of the jobs through the system are known together with the distribution for the duration and the number of jobs. This is all the data needed for the model and thus the coding can be done.

The coding can be done in many different software programs. Think of Simul8, AnyLogic, SimScale or Tecnomatix Plant Simulation (TPS). For this thesis TPS will be used, as this software is known to me.

Nature of simulation models and simulation output

Simulation models can be terminating or non-terminating simulations . Terminating simulations have a natural endpoint. This can be the closing time of a shop or the end of the time-period of the investigation or the completion of a schedule. . Non-terminating simulations will never stop. "There is no reason for the simulation to stop other than the user interrupting the run.". The length of the simulation has to be determined by the user for these kind of simulations.

The department of OBI is a 24/7 active control centre. This means that the simulation should be a non-terminating simulation.

Non-terminating simulations often reach a steady-state . Steady-state means that the output of the KPI's will vary according to some sort of a distribution, called the steady-state distribution. This steady-state can be two forms. It can be a steady state point/line, meaning that after reaching this steady-state the output will vary according to the point/line. Another option is the steady-state cycle. This means that there is a certain pattern that keeps repeating itself.

Warm-up length

For the validity of the output it is important to know the warm-up length. This can be determined by making use of the Marginal Standard Error Rule (MSER) or a graphical method (same as replications graphical method). “MSER has the aim to minimise the width of the confidence interval about the mean of the simulation output data.” .

$$MSER(d) = \frac{1}{(m-d)\cdot 2}\sum_{i=d+1}^{m}\left(Y_i - \bar{Y}(m,d)\right)^2$$

d = the proposed warm-up period
m = the number of observations in the time-series of output data
$\bar{Y}(m,d)$ = the mean of the observations from Y_{d+1} to Y_m

For every *d*, the MSER will be calculated. This means that for every *d,* the difference between the current mean and the mean of all the data coming up later is determined. The *d* which has the lowest MSER will be the warm-up period. Reason for this is that the difference of the current KPI value and the value in the future of the simulation is the lowest.
There is one constraint to the determination of the warm-up period. If the warm-up period is more than half of the run length used to determine the MSER, this value is not accountable and the MSER has to be determined out of a larger data sample.

Replications

Not only the warm-up period is important to know, also the length of the run and the amount of replications of the run are important. Running the simulation too short can cause that the data is too little to make proper analysis out of. However, running the simulation very long increases the run-time. The same counts for the number of replications. The more replications, the more data there is, but the runtime increases as well. Therefore, for both terms, the distinction has to be made between runtime and the amount of data.

“A replication is a run of a simulation model that uses specific streams of random numbers.” . By changing the streams of random numbers, more randomness is inserted in the model, resulting in a more accurate output.

Robinson (2014) describes three approaches to come up with the amount of replications, namely: a rule of thumb, a graphical method and a confidence interval method.

A rule of thumb

Robinson (2014) states that according to Law and McComas (1990) at least three to five replications are needed. This method does not look into the output of a model, but is just an advice that could be followed.

Graphical method

The graphical method uses the cumulative mean of multiple replications to determine the amount of replications needed. The first step of this method is to make a lot of replications of a simulation model (around 20). For every of these replications, the runtime is the same and the warm-up period is used. Every replication gives a KPI value. A graph can be made with all cumulative means, one mean for every amount of replications. The point at which the graph is nearly flat, is the amount of replications needed.

Confidence interval method

This method is an addition to the graphical method. Instead of plotting the cumulative means, a confidence interval is made. The goal of this method is to achieve a deviation percentage lower than

5%. In the book of Robinson (2014), this method is explained in detail. This method uses formulas to determine the actual deviation for every number of replications.

Because of the 10 week time constraint and because of the little difference between the confidence interval method and the graphical method, it is chosen to use the graphical method for both the replications as well as the warm-up length. This method is a lot quicker to determine and almost as accurate as the confidence interval method/MSER. Because of the decision to use the graphical method, the confidence interval method is only shortly described, as it is of no usage for this research.

Run length

After determining the number of replications, the run-length should be determined. There is almost no information on how to calculate the run-length. However, according to Banks et al. (2009), the run-length should at least be 10 times the warm-up period. Therefore this will be used in the thesis.

, the distributions will be determined by following the 4 steps. However, some of the datasets are not compatible for the analysis, so extra coding is needed. After that, some of the data has to be removed. For example calls with a duration of 0 seconds. When the datasets are ready, the analysis starts.

What is noticeable is that the duration of calls as well as the amount of calls is very fluctuating. If the dataset is divided into the 3 shifts, the variability will be too large, resulting in an unrepresentative sample. Therefore, the decision has been made to divide the data into hours. For every hour, the distribution will be determined. This results in less variability. As can be seen in ***Error! Reference source not found.***, by adding extra shifts, the averages lay closer to the actual number of calls.

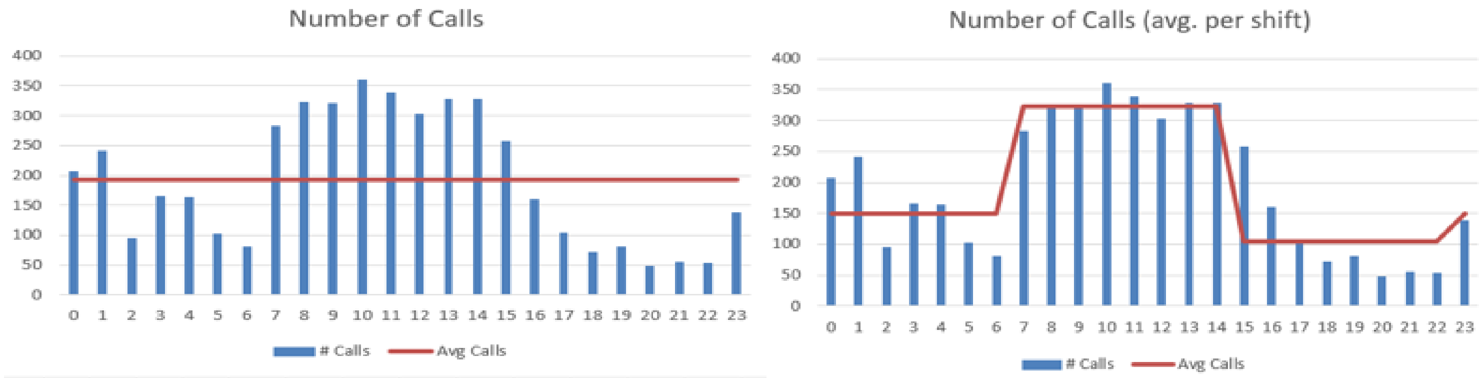

Figure 7: Averages of #calls

In *Figure 7*: Averages of #calls, the calls are divided into 3 groups of 8 hours. What can be seen is that the average is still far off from the actual calls at some hours. Therefore the decision is made to divide into 24 groups of 1 hour. This means that the averages are more closely related to the real world, however this also increases the amount of distribution fitting activities. Instead of 24 times, now 72 times it has to be determined which distribution fits best.

The number of sequences per day is very dependent on what day it is. Per day a different amount of maintenance jobs are present. So the decision is made to find distributions for every day of the week.

The duration of checking sequences is not known. This duration cannot be found via the logbooks, nor can it be observed, since the number of days to observe is too large to be able to do it within the 10 week time frame. Therefore the decision is made to use an experts opinion as data source. This expert is asked what the average duration is of checking switch sequences. He told that the checking could differ from 15 minutes to 1 hour, with an average around 25 minutes.

The distributions skewness is positive, which means aimed to the right. (Jain, 2018). Therefore the potential distribution which could fit are: gamma, Weibull, exponential and lognormal distribution.

In discussion with the employees of the department, they thought that the gamma distribution is the closest in their opinion to the reality. So the gamma distribution is chosen. However, the gamma distribution has 2 parameters. As it is unable to determine these parameters statistically, it is done by trial-and-error (plotting the graph for different parameters and choose closest). During this trial-and-error period, different variance levels have been used. By using the expert opinion, it was concluded that the number of times the duration was near the average of 25 minutes. Therefore the decision is made to use a gamma distribution with little variance.

4.5 Results

The results of the distribution can be found in *Appendix C.* Since the number of graphs is too much to put in the appendix, only one solution is given. To be able to understand the solution, below an explanation will be given on what parameters are used for every distribution.

The Gamma distribution is declared with the rate and scale parameters. The rate parameter is noted with the letter α, the scale parameter with the β. The mean can be calculated by doing $\frac{\alpha}{\beta}$ and the variance by using the formula $\frac{\alpha}{\beta^2}$.

The lognormal distribution is declared with the parameters (μ, σ^2). The mean of the distribution can be calculated by using the formula $e^{\left(\mu+\frac{\sigma^2}{2}\right)}$. The variance can be calculated using $e^{\sigma^2}e^{(2\mu+\sigma^2)}$.

The Weibull distribution is declared using the scale parameter λ and the shape parameter κ. The mean can be calculated using the gamma distribution in the following way: $\lambda\Gamma(1+\frac{1}{k})$. The variance can be calculated using $\lambda^2[\Gamma\left(1+\frac{2}{k}\right)-\left(\Gamma\left(1+\frac{1}{k}\right)\right)^2]$.

The actual parameters which are used in the simulation model can be found in *Appendix C*: Data analysis. In the summary, per type of data, the way of coming up with the variables can be found.

4.6 Summary

To sum up this chapter, per type of work the distribution will be given with a summary of the explanation.

Sequences
The number of sequences is determined by a Poisson distribution. The mean for this distribution is determined by analysing the data. For every day the average is calculated and used. The duration of performing the sequence is also done by data analysis. The duration of checking the sequences however is not in the data. Therefore expert opinion is used, as earlier said. The values of the variables can be found in Appendix F.

Failures
The number of failures is determined by expert opinion. Since the number of failures differs every week, there is no standard. However, due to the risk of not being able to handle the failures, it is chosen in discussion with the experts that it is most useful to add exact one failure every hour. This reduces the risk of not being able to handle the jobs in time due to an unexpected failure. The

duration of a failure is also determined by expert opinion. Due to the fact that the failures cannot be checked in the system. The average is around 25 minutes as well, as can be seen in Appendix F.

Calls

The number of calls and the duration of calls is determined by data analysis. The call logbook is checked and for every hour a distribution for the duration is given. The number of calls is determined by a Poisson distribution with the average amount of calls per hour as variable.

Manually switching

For the manually switching it is not possible to check data or ask experts, since it differs that much every day. The experts told that it was best to keep it in mind and whenever the utilization is around 0.92, think of an extra employee to be added.

Chapter 5: Solution design

This chapter describes the models used. In section 5.1 the mathematical model used is explained. In section 5.2 the simulation model is described.

5.1 Mathematical model

To determine the number of employees needed per shift, an integer linear programming model will be used. As explained in *4.4 4.3 Determination of* distributions

Within the domain of statistics, there are many ways to determine distributions. For this thesis, only one of these methods is used to determine whether a distribution can be found for the data.

To determine the distributions, a statistical programme or software will be used to easily calculate and plot the data and distributions. This software will be R. At the website of R , the programme is described as: “R is a system for statistical computation and graphics. It provides, among other things, a programming language, high level graphics, interfaces to other languages and debugging facilities.”

In R, the package ‘fitdistrplus’ will be used. This package is made specifically to determine distributions of datasets **.**

The method to determine the distributions, will be the same as described in the paper of M.L. Delginette-Muller et al. (2009). This method consist of the following steps:

5. Graphical display of the dataset
6. Characterization of the dataset
7. Fitting of a distribution to the dataset
8. Simulation of the uncertainty of the distribution to the dataset

The above steps is the general way of finding a distribution. However, it is not clear yet, what to do at each step. Therefore, per step, a detailed explanation will be given on how every step will be performed. Within the explanation, the small lines of code used in R will be stated.

Step 1: Graphical display of the dataset

The graphical display of the dataset will be given by plotting a histogram as well as plotting the cumulative distribution of the data. These graphs will give an insight on how the data is divided. This can give already some ideas of which distribution might fit to the data.

Step 2: Characterization of the dataset

In this step, the descriptive statistics of the dataset will be determined. This includes the: minimum, maximum, median, mean, standard deviation, skewness and the kurtosis. Next to this, the Cullen and Grey graph will be plotted. This graphs shows, by plotting the square of the skewness and the kurtosis, potential distributions that lie close to the dataset.

Step 3: Fitting of a distribution to the dataset

This step is closely related with step 2. In step 2, the Cullen and Grey graph is plotted. As said, this gives potential distributions. In this step, in order of closest related to less related, the distributions will be fitted to the data. This means that the data will be plotted together with the distribution in one graph. Next to this graph, the QQ-plot, PP-plot and cumulative distribution functions will be shown. All these graphs have the objective to show whether the distribution is representative for the dataset. Furthermore the rate and the shape will be determined.

If a distribution is closely related on sight, a goodness-to-fit test will be used. For continuous distributions, the Kolmogorov-Smirnov test will be used. For discrete distributions, the chi-square test can be used. The Kolmogorov-Smirnov test is as follows in R:

Next to the test, also the specific parameters of the distribution will be determined. This is done with the following code:

During the Kolmogorov-Smirnov test a p-value will be calculated. If this p-value is larger than 0.05, the H_0 hypothesis cannot be rejected and thus the chosen distribution represents the dataset in a good enough way.

Step 4: Simulation of the uncertainty of the distribution to the dataset

The last step is about bootstrapping the dataset. This gives per data point the confidence interval it can lie in and also plots the bootstrapped values. The summary gives the median as well as the 95% confidence interval of the dataset.

Conclusion

To conclude, step 1 to 3 determine the distribution which will represent the dataset. If the p-value is larger than 0.05, the distribution can be used as representation. Furthermore, by doing step 4, a confidence interval can be made of the parameters. This gives, together with the graphs, a good overview of how the variability is, when random generating values multiple times. This can later be used to adjust the parameters to more extreme values (more work, so less risk of under capacity).

4.4 Determination of distribution, the number of employees needed will be calculated per hour. Furthermore, per day the calculation will be done as well. This is due to the fact that input for the model (e.g. duration of calls) can differ a lot per day. Next to this, the model will be executed for both the North and South station separately and the scenario when combining them.

Below the model will be described. First of all the explanation is given on what indices are used in the model. After that the optimization rule and the constraints are given.

Description of the Linear Program

This section will give the letters with their explanation. Next to this, all the possible values that a parameter can have, will be given.

t
- *1, handling of the incoming calls*
- *2, executing the switch sequences*
- *3, checking the switch sequences*
- *4, handling the disruptions*

d
- *1, Monday*
- *2, Tuesday*
- …
- *7, Sunday*

h
- *0, 00:00 – 01:00*
- *1, 01:00 – 02:00*
- …
- *23, 23:00 – 00:00*

Parameters

N_{dht}	*Number of tasks **t** occurring during hour **h** on day **d** where d = 1, 2,..,7, t = 1,2,3,4 and h=0,1,..,23.*
T_t	*The duration of one task **t** where **t** = 1,2,3,4*

Decision Variables

E_{dh}	*Number of employees starting their shift on day **d** during hour **h** where d = 1, 2,..7 and h=0,1,..,23.*

Optimization rule

The goal of this study is to minimize the number of employees per week. This is described in the following mathematical rule.

$$Min \sum_{d=1}^{7} \sum_{h=0}^{23} E_{dh}$$

Constraints

Constraints are added to the model to improve the validity according to the reality. Furthermore, some constraints have to be added, because that is a rule of the company. First of all the constraints will be stated, later on the explanation is given.

$$E_{dh} \geq 1 \qquad \forall d, h \in \{7,15,23\} \qquad (1)$$

$$\sum_t (\sum_{h=7}^{14} (N_{dht} * T_t)) \leq 3600 * 8 * E_{d7} \qquad \forall d \qquad (2)$$

$$\sum_t (\sum_{h=15}^{23} (N_{dht} * T_t)) \leq 3600 * 8 * E_{d15} \qquad \forall d \qquad (3)$$

$$\sum_t (\sum_{h=0}^{6} (N_{dht} * T_t) + (N_{d23t} * T_t))) \leq 3600 * 8 * E_{d23} \qquad \forall d \qquad (4)$$

Constraint (1) is the constraint that says that at least one employee should be present at the stations. This is a legal rule. At any time there should be a person that can switch off the electricity network in case of emergency. As only the employees are allowed to switch, there has to be at least one employee at the station. Constraints (2 - 9) are constraints that make sure that there are enough employees to finish the work within the hour. As employees work in shifts of 8 hours, every employee that started 7 or less hours before the concerning hour will be present at the stations and therefore works. An hour has 3600 seconds, therefore it is multiplied with the amount of employees during that hour of the day.

For the readability of this thesis, all the assumptions will be stated below. It can occur that some of them have been mentioned before.

- The sequences can only be checked between 23:00 – 01:00. So all the sequences in a night are equally divided over the 2 hours.
- The first execution of the sequence will be done in the same hour as within the checking happened. The second execution will all be done in between 05:00 – 06:00.
- Failures can only occur from 7:00 hour until 23:00, because of the lack of trains in the other hours of the day. In every of those hours 1 failure will occur in the model (chapter 4). This assumption is made to make sure that at every hour the department is able to handle all work including a failure. Since failures occur very randomly and they have to be handled quickly, because they affect travellers a lot, it may not happen that too little employees are scheduled whenever a failure occurs. Therefore one failure is taken into account for every hour.
- The amount of calls per hour as well as the duration of them is determined by the upper boundary of the 99% confidence interval of the dataset. This is done to make sure that the risk of under capacity is decreased. Under capacity can raise far larger problems, which can possibly affect the traveller, which may not happen.
- Every employee works at the same speed.

Next to the assumptions to the model, there are restrictions which have to be taken into account for in the mathematical model. The shifts of the employees start at 7:00, 15:00 or 23:00. This means that the following addition can be made to the model.

$$\mathbf{E_{dh}} \quad \begin{cases} 0, & \forall h \in \{0,1,..,6,8,9,..,14,16,17,..,22\} \\ > 0, & \forall h \in \{7,15,23\} \end{cases}$$

5.2 Simulation model

As said in *Chapter 2: Literature review*, a simulation model will be made in addition to the mathematical model to be able to calculate some of the KPI's as well as to verify whether the solution found by the mathematical is a good solution. The simulation model will be made 3 times. The first scenario is for the North station, the second scenario is for the South station and the third and last scenario is for the combination of the stations. Every model will have different input, however, the flow of the jobs through the system will be the same. All models have the same flowchart, which can be seen in *Figure 5*: Flowchart of the system.

This simulation model will give a basic overview of the stations. To let the model run properly, some extra assumptions have to be made. These extra assumption(s) will be stated below.

- The jobs are handled according to the "First-In First-Out" method (FIFO). No matter what type of jobs are waiting.
- Every hour, the input distributions will change according to one of the tables in *Appendix D:* Solution distribution fit
- Whenever an employee is working on a job, no matter of the type of job, he cannot work on another job until the job is finished.

Input

The input for the simulation model is the input for the mathematical model combined with the output of the mathematical model. This means that all distributions which are used in the mathematical model are also used in the simulation model. However, some of the averages which are determined, could not be used without a distributions. Therefore, it is chosen that all input regarding the number of calls, failures and sequences is from a Poisson distribution, with the mean as lambda. The table for the amount of calls and sequences can be found in the Appendix F.

Output

The output of the simulation model is the average waiting time of the jobs. This is the only output needed, since the number of employees is known at the start and the utilization of the employees is combined within the waiting time of the jobs. The lower the waiting time, the lower the utilization.

The validation of the simulation will be done by checking the behaviour of the model. By changing the number of employees up and down, the difference in KPI can be measured. By comparing the measurement with the expectation, the behaviour can be controlled. If the behaviour is according to the expectation, the model is valid and the different scenarios can be checked. Whenever the solutions will be given, the model will be tested first whether it reacts as expected. This can also be seen in the results.

Warm-up period

The warm-up period is determined by making use of the graphical method. By plotting the value of the KPI per day, a graph is made which can be seen in *figure 9*.

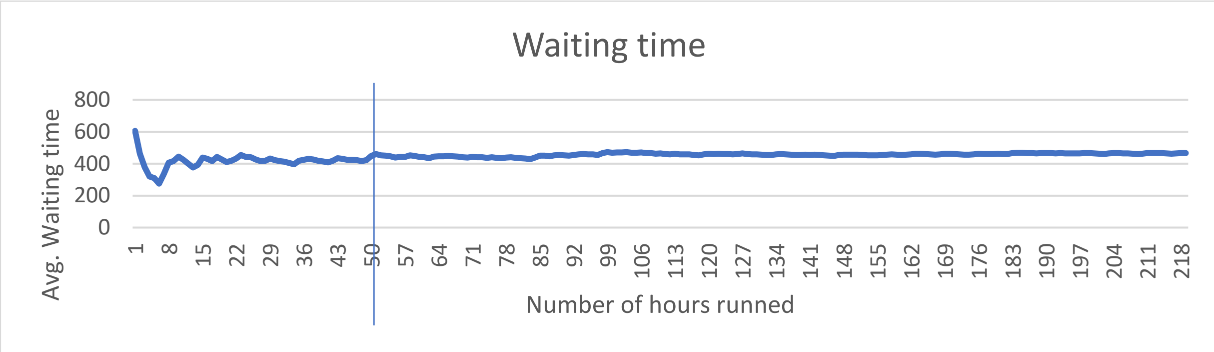

Figure 8: Warm-up determination

This graph shows the fluctuation in the KPI. What can be seen is that the KPI is instable until around day 50. After day 50 the KPI is stable for some weeks, after a small jump can be seen again. However, the warm-up period has been set on 55 days. The reason to do 55 days, is because this gives some margin on the exact point of warm-up period. By adding 5 days to the 50 days, there is a small safety buffer added. To choose not for an even longer warm-up period, is because of the difference in values of the KPI is not that much. By increasing the warm-up period even more, the run time will be very long. The rule for the run length is at least 10x the warm-up period. For a warm-up of 55 days this means 550 days run length. By increasing the warm-up length to for example 105 days, the run length goes to 1050 days, which is very long and takes loads of time. Therefore the decision is at 55 days warm-up.

Replications

The number of replications will be determined by using the graphical method. First of all, 10 replications have been made, by running the simulation model 10 times, all with different seed values. By writing down one of the output variables and determining the cumulative mean, it can be seen from figure 10 how the mean differs between the replications. Whenever the graph goes flat, that number of replications is chosen, since more replications do not change the average effectively.

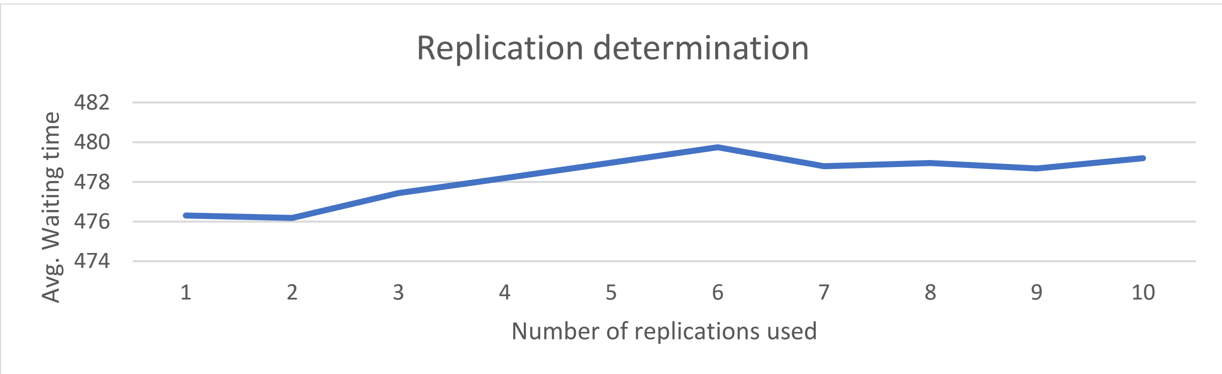

Figure 9: Replication determination

In this case the replication size is 3 times. When looking at the graph it can be seen that the average fluctuates around 2 seconds. So as little replications as needed are used. According to the rule of thumb (Robinson, 2014), the minimum amount of replications to make is 3. Since the difference in waiting time between 3 replications and more is less than 2 seconds, 3 replications will be used.

Run length

The run length is as earlier said equal to at least 10x the warm-up period. Since in this case the difference between 10x or more times is little, 10x the warm-up period is chosen as run length. So in this case the run length is 550 days (Robinson, 2014).

Chapter 6: Results

This chapter shows the results of the mathematical model in chapter 6.1, the results of the simulation model in chapter 6.2 and the utilization levels in chapter 6.3.

6.1 Mathematical model

The solving of the mathematical model is done for every station. By using the Excel solver, the mathematical model can be solved. The results can be seen in *Figure 11*, *Figure 12* and *Figure 13*.

# Employees	**Hour**	23	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Day	***NORTH***	Night								Morning								Evening							
Monday		2								2								1							
Tuesday		2								2								1							
Wednesday		2								2								1							
Thursday		2								2								1							
Friday		2								2								1							
Saturday		1								1								1							
Sunday		2								1								1							

Figure 10: Solution North

# Employees	**Hour**	23	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Day	***SOUTH***	Night								Morning								Evening							
Monday		1								1								1							
Tuesday		1								1								1							
Wednesday		1								1								1							
Thursday		1								1								1							
Friday		1								1								1							
Saturday		1								1								1							
Sunday		1								1								1							

Figure 11: Solution South

# Employees	**Hour**	23	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Day	***COMBI***	Night								Morning								Evening							
Monday		2								3								2							
Tuesday		2								3								2							
Wednesday		3								3								2							
Thursday		2								3								2							
Friday		2								3								2							
Saturday		2								2								2							
Sunday		2								2								2							

Figure 12: Solution Combi

By comparing these values with the current situation, which can be seen in *Table 2,* it can be seen that for the South station, there is no difference, meaning that that is already the most efficient way of scheduling. However, the North station does give a difference. Instead of the current situation with 2 employees every shift, the mathematical model gave 1 employee every evening shift. However, with the in-availability of the manual shifting data as stated in chapter 4, this solution has to be checked using the simulation model.

6.2 Simulation model

The simulation model is solved with regard to the replications, warm-up length and run length. For all the scenario's, first of all the old scenario is performed. After that, the 'new' scenario is performed and the solutions are given.

The validation of the model is done by looking into the results of the model. Decreasing the number of employees should increase the waiting time per job. The results in Appendix E show that this is indeed true, meaning the model reacts in the proper way.

North station

The current scenario for the North station is 2 employees at every shift. By inserting this number of employees into the simulation model, the KPI's are calculated. The results of this can be seen in table 3 and in *Appendix E.*

Table 2: Solution North current situation

KPI	Value
Number of employees per week	42
Average waiting time of a job (seconds)	219,11

What is noticeable is that the waiting time is high for the system. However, this has multiple reasons. First of all, if more than two sequences appear in one hour, at least one should wait if appearing at the same time. This means that the waiting time of that sequence is around 25 minutes. Next to that waiting sequence, also the calls arriving then have to wait, meaning a waiting time of around 25 minutes as well. If this happens often, the waiting time will increase quickly. To emphasise this behaviour, a screenshot is made of the average waiting time in seconds per job type, see *figure 14*. What can be seen is that the waiting time for the sequences is high and affecting the calls a lot.

```
avgWT=218.493499098044
avgWTseq=229.082042867935
avgWTfail=70.5858946123698
avgWTcall=254.030898609617
```

Figure 13: One solution North current scenario

To be able to validate the solutions, the difference in waiting time will be determined and analysed. As the waiting time is not according to the reality, but the waiting time behaves at it should do, the difference in waiting time should be the actual difference in real life when changing the number of employees per shift.

So to be able to analyse the difference, the 'new' scenario is simulated as well. By inserting the number of employees according to the mathematical output, the model was simulated. The results can be found in table 4 and in *Appendix E*.

Table 3: Solution North 'new' situation

KPI	Value
Number of employees per week	32
Average waiting time of a job (seconds)	559,65

When looking in the differences of the number of employees per week, 10 employees are scheduled less. Especially during the evening shift, 1 employee is scheduled less. The effect of those 10 employees less is that the average waiting time is increased with around 340 seconds. This is very much, however, the solution is a very influenced by the number of failures in the system. Looking at *figure 15,* shows that the average waiting time of failures is increased with 500 seconds in comparison with the current scenario. The reason for this increasement is the assumption made on forehand. It is assumed that every hour one failure occurs. With 2 employees, this assumption can be handled fairly quickly, however with only 1 employee, this assumption is highly influencing the system, since this employee only has 35 minutes of the 60 minutes to solve something else than a failure.

```
avgWT=563.992317770213
avgWTseq=346.111228435697
avgWTfail=572.176096611368
avgWTcall=511.742760254133
```

Figure 14: One solution North 'new' scenario

To be able to see the influence of the failures, the simulation is ran again but with half the amount of failures. The results can be seen in *figure 16.* The waiting time of the failures is decreased with 130 seconds on average and the total average waiting time even decreased with around 170 seconds.

```
avgWT=395.702929046949
avgWTseq=222.971902226181
avgWTfail=444.35875425326
avgWTcall=422.808146190188
```

Figure 15: One solution North 'new' scenario with 0.5 failures

South station

For the south station, there is no need to simulate this station separately. Since the mathematical model already shows that it is possible with 1 employee for all the shifts and this is also the current scenario, there is no need to simulate this. The solution between the 'new' scenario and the current scenario will be the same since the settings are the same for both.

Combination station

The current scenario for this station will be the amount of employees on both stations added. This means that for every shift there are 3 employees. The results are shown in table 5 and *Appendix E.*

Table 4: Solution Combi current situation

KPI	Value
Number of employees per week	63
Average waiting time of a job (seconds)	1627,14

For the Combination station, the same counts as for the North station. The focus will be on the difference in waiting time, not on the waiting time itself.

The same model is ran another time, but with the new schedule, determined with the mathematical model. The output can be found below in table 6 and *Appendix E.*

Table 5: Solution Combi 'new' situation

KPI	Value
Number of employees per week	48

Average waiting time of a job (seconds)	2749,95

The difference between the KPI's is very large, about 1125 seconds on average. To look where the difference is, two screenshots of the output of the simulation model are shown in *figure 17.*

What can be noticed is that the effect is little and the difference is still big. This means that the difference is made somewhere else. When looking at the waiting times for the jobs separately, it is noticed that the waiting time for the sequences is very large. This means that the effect of decreasing the number of employees in the night shift is large. By running the same model again but now with 3 employees in the night shift instead of 2 the results of figure 18 occur.

avgWT=2748.23712327493
avgWTseq=3609.29850680885
avgWTfail=2275.52723091609
avgWTcall=3142.80316233598

avgWT=1630.7341655416
avgWTseq=2196.30348671354
avgWTfail=1069.82892486275
avgWTcall=1947.44099619777

Figure 16: Solution Combi new (left) vs current (right)

The difference can mainly be explained by the difference in failure and sequence handling. The sequence waiting time increases with 1400 seconds and the failure waiting time with 1100 seconds.

By increasing the employees working the night shift to 3 and running the model again, it can be checked whether this would work for the system. The results can be seen in *figure 18*.

avgWT=2051.80191295124
avgWTseq=2184.37699683484
avgWTfail=2112.87977932919
avgWTcall=2417.8297078396

Figure 17: Solution Combi with 3 employees at Night

The waiting time for the sequences is at the same level as with the current situation now, however the difference in the waiting time of failures is still that high. To counter this increasing value, the number of employees working the evening shift should be increased. By doing this, almost the same input is used as at the current situation, which is no improvement.

The actual targets of the KPI's will be mentioned in chapter 7, as well as, the usefulness of the new model compared to targets and the current system.

6.3 Utilization results

Next to the results of the mathematical model and the simulation model, it is useful to get an insight in the utilization level for the different stations if the 'new' scenario is used and how it is in the current situation. These utilization levels can be found in the figures below. In the next chapter, the utilization will be used to come up with an advice for the company.

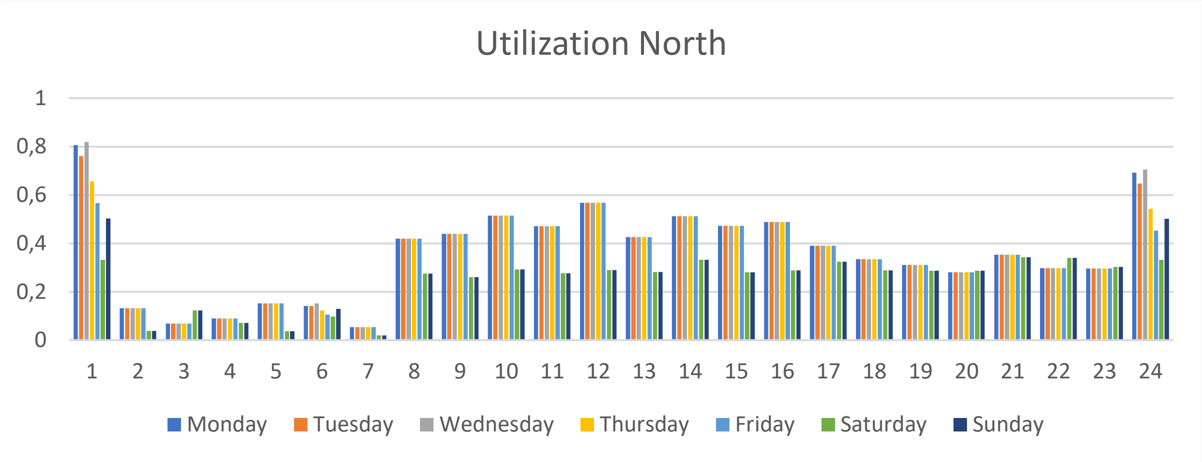

Figure 18: Utilization level North current

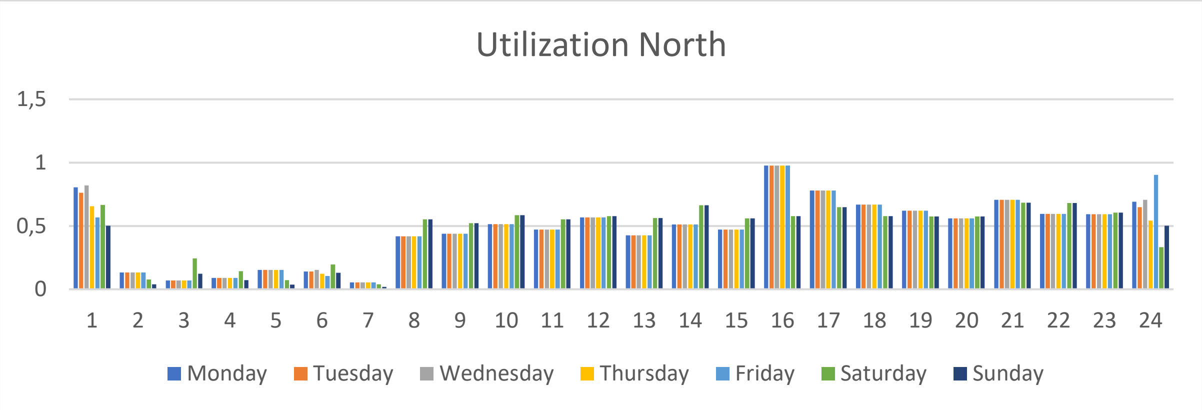

Figure 19: Utilization level North 'new'

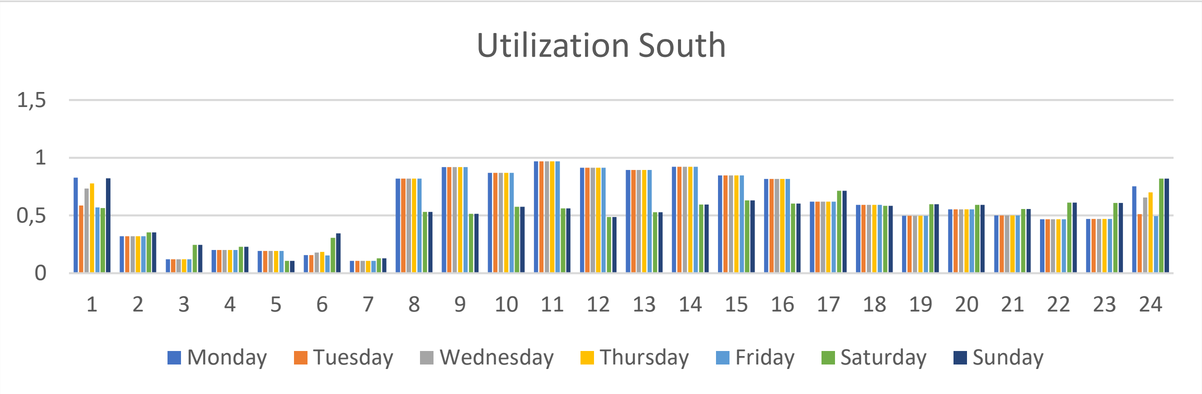

Figure 20: Utilization level South current & 'new'

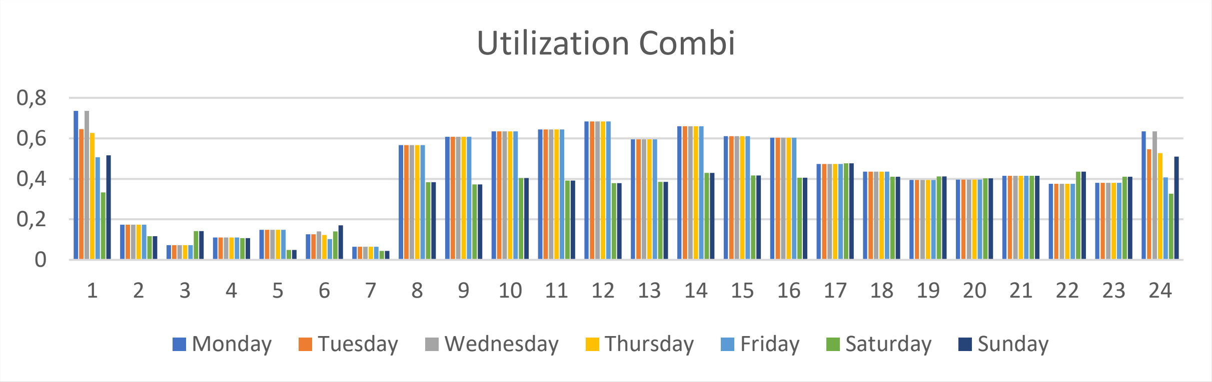

Figure 21: Utilization level Combi current

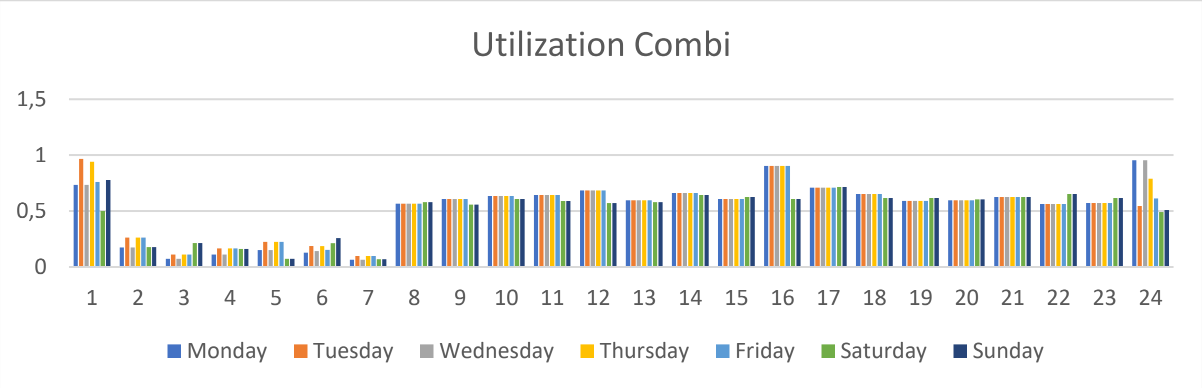

Figure 22: Utilization level Combi 'new'

The results of these utilization levels will be concluded in the next chapter. Also an advice and recommendations will be given there, according to the difference in utilization between the current and the new system.

Chapter 7: Conclusions and advice

This chapter will draw the conclusions in chapter 7.1, will give an advice to the company in chapter 7.2 and discusses the results and the advice in chapter 7.3.

7.1 Conclusions

In this chapter the conclusions will be drawn out of *Chapter 6: Results.* However, before doing so, a short sum up of the approach will be given. This research uses the method of first making a mathematical model and solve it, after than use a simulation to verify the solutions of the mathematical model.

In chapter 6 already some conclusions were made. These conclusions will be summed up here. These conclusions will be given per separate station. In the advice the conclusions will be summed up.

North station

The result of chapter 6 show that by reducing the number of employees, the waiting time increases with around 6 minutes on average. The targeted time of handling the jobs is 3 minutes. This means that the increasement in waiting time is not according to the target. This difference is all because the amount of employees on the evening shift is reduced to 1. As stated in the chapter, the main reason for the increasing waiting time is the failure rate. Since there is 1 employee less to focus on the failures, these failures increase the waiting time of both the calls and the failures drastically.

The utilization for the new scenario can be seen in the figures in chapter 6.3. What can be seen is that for some hours, the utilization is high whereas for some hours it is very low. The hours 00:00-01:00, and 15:00 – 16:00 have a very high utilization of around 0.9. This means that for these hours, it can occur that there is under-capacity, since the amount of work for the handling of the extra manual switching is not in the model.

South station

For the south station the conclusion of the calculations are easy. The amount of employees currently at the station is already the minimum amount of employees that are needed. The conclusion can be drawn that there are enough employees on that station.

The utilization level for the South station is high. However, it can be managed with one employee as this is possible in reality right now.

Combination station

The results for this station show that by decreasing the amount of employees for the night shift, the waiting time per job increases a lot. This is due to a big difference in the waiting time for the sequences. By adding an extra person for all the night shifts, the difference is far less between the current and the new scenario, but still too big. Therefore some extra employees should be scheduled in the evening shift, resulting in almost no improvement on the number of employees scheduled per shift. However, with the addition of the manual switch work, the variability increases. The more variability the more beneficial it is to combine the stations, since the combination of two variances results in a lower variance than their addition.

7.2 Advice

My advice for ProRail and the OBI department will combine all the station conclusion into one advice. First the actual advice will be stated, after which the arguments will be given.

- Keep the stations separated, set the number of employees working at North during the evening shift to 1, but add an extra flexible employee working other times than the normal shifts and able to work on both North and South station.

The reason to keep the stations separated is because the decrease in number of employees is little compared with the new solution for the North station and the South combined. One of the problems of combining the stations is that the employees have to do a cursus for the other part of the Netherlands (either North or South). This cursus costs money. Since there is a solution that is almost as good as the combination, but does not need extra cursus, this is both an improvement for the employees as well as the company. The employee does not have to do a cursus and the company does not have to pay.

But why decreasing the number of employees working the evening shift at the North station to 1 and add a flexible employee? The reason to decrease the number of employees in the evening shift is because the increasing waiting time of the jobs is little compared to the current solution. Furthermore, this increasing waiting time is for a very large part based on the failure assumption in the simulation model. As explained earlier, this failure assumption increases the waiting time a lot. Therefore, when looking at the mathematical model as well as the simulation model, in my opinion the company could decrease the number of employees working on the North station in the evening shift. The reason to add the extra flexible employee is because the utilisation of both stations is quite high during the morning and evening shifts. Especially the hour from 16:00-17:00. By adding an employee which works from 09:00-17:00 and can work on both stations, the peaks can be managed more easily. This means that all employees work effectively during their shifts. The simulation results of this advice can also be seen in appendix E. The increasement of waiting time is around 3 minutes, which is within the target time.

The addition of the extra flexible employee, will not increase the number of employees needed per week, but will increase the flexibility of the workforce scheduled. In the most busy hours there are more employees, and in the quieter hours, less employees are scheduled. Also the work pressure on the employees might decrease, which is both beneficial for the employees as well as the company. Employees that feel less pressure often has more fun in its job.

The combination station will decrease the amount of variability in the system. Since the unknown time of the manual switch handlings, there is even more variability in the North and South stations. By combining them, the variability decreases. However, it is to the company to think of whether the amount of variability is large enough to really benefit from it. In this case, the potential benefit of combining the stations is little, therefore the decision to add a flexible worker would be more beneficial. This is a little step towards the building of the combination station.

7.3 Discussion

In this research, the assumption is made that every hour the employees should be able to handle at least 1 failure. This assumption has a lot of consequences, both pros and cons. As pro, it can be said that by adding this assumption, there is almost no chance of under capacity. Under capacity for this department is horrible, so the decision to add this assumption helps in reducing the risk.

From the other side, it can be said that there is a possibility that the solutions of this research are too high, since almost no failures occur for example. If no failures occur, way too many work is counted in comparison with the actual work.

So for the next study towards this case, the best method would be to try to gather as much data as possible and use statistics on it. For this case, the 10 week time frame denied that option, but with a longer time frame this might work.

Another point of discussion is the lack of information about the sequence checking. In this thesis this information is added via expert opinion, however, the variability is very hard to determine in parameters. The experts tell the information, but the change into a parameter is almost impossible. This causes the models to be less realistic, however, the difference between the reality and the expert opinion will not differ so much to really change the outcomes of this research.

Future projects that could be done on the OBI is looking into possibilities of automation, for example with sequence handling. Also on the other departments of ProRail there are some projects that could be done. For example try to think of a way to schedule the maintenance differently, reducing the impact on the train traffic as much as possible. Or think of ways to insert the data out of the phone calls into the system automatically.

Bibliography

BusinessDictionary. (2019). *deterministic model*. Opgehaald van BusinessDictionary: http://www.businessdictionary.com/definition/deterministic-model.html

Heerkens, J., & Winden, A. v. (2012). *Geen probleem, een aanpak voor alle bedrijfskundige vragen en mysteries.* Buren: Business School Nederland.

Infrabeschikbaarheid ProRail. (2019). *Infrabeschikbaarheid*. Opgehaald van ProRail Sharepoint: https://prorailbv.sharepoint.com/sites/info-over-assetmanagement/SitePages/Infrabeschikbaarheid.aspx?web=1

Jain, D. (2018, August 23). *Skew and Kurtosis: 2 Important Statistics terms you need to know in Data Science*. Opgehaald van Codeburst: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa?gi=fa5c37b78da3

Marion, G., & Lawson, D. (2008). *An Introduction to Mathematical Modelling.* Bioinformatics and Statistics Scotland.

Morris, T. P., White, I. R., & Crowther, M. J. (2017, December 8). *Using simulation studies to evaluate statistical methods*. Opgehaald van Cornell University: https://arxiv.org/abs/1712.03198

Pinedo, M. L. (2009). *Planning and Scheduling in Manufacturing and Services.* New York: Springer Science + Business Media, LLC.

ProRail. (2019). *Over ProRail*. Opgehaald van ProRail: https://www.prorail.nl/reizigers/over-prorail

R Core Team. (2019, April 26). *R Language Definition*. Opgehaald van R-Project: https://cran.r-project.org/doc/manuals/r-release/R-lang.html

Robinson, S. (2014). *Simulation: the practice of model development and use.* London: Macmillan Education Uk.

Stephanie. (2016, November 3). *Stochastic Model/ Process: Definition and Examples*. Opgehaald van Statistics How To: https://www.statisticshowto.datasciencecentral.com/stochastic-model/

Appendix A: Organization chart

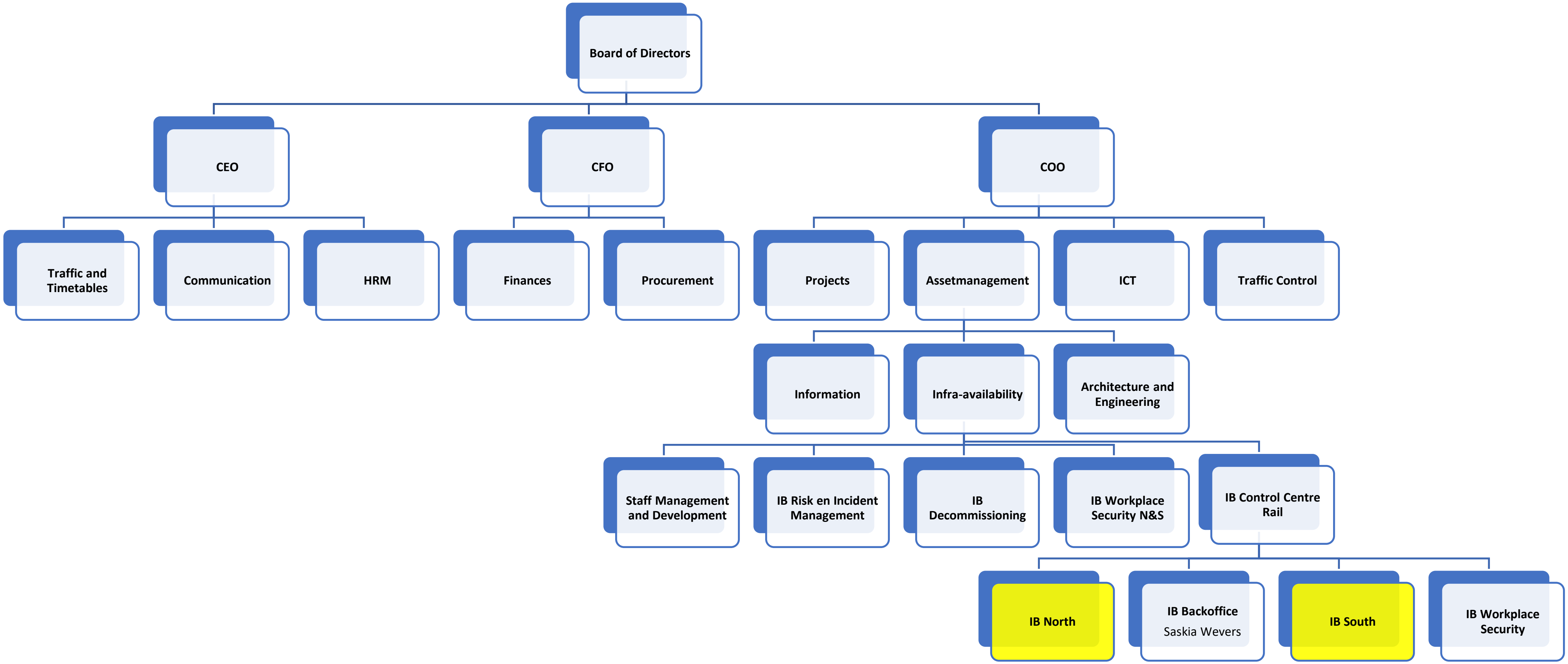

Figure 23: Organization chart

Appendix B: Problem cluster

This appendix shows the problem cluster. The arrows stand for causal relationships. Some of the problems cannot be fixed, therefore a legend is added to distinct the different types of problems.

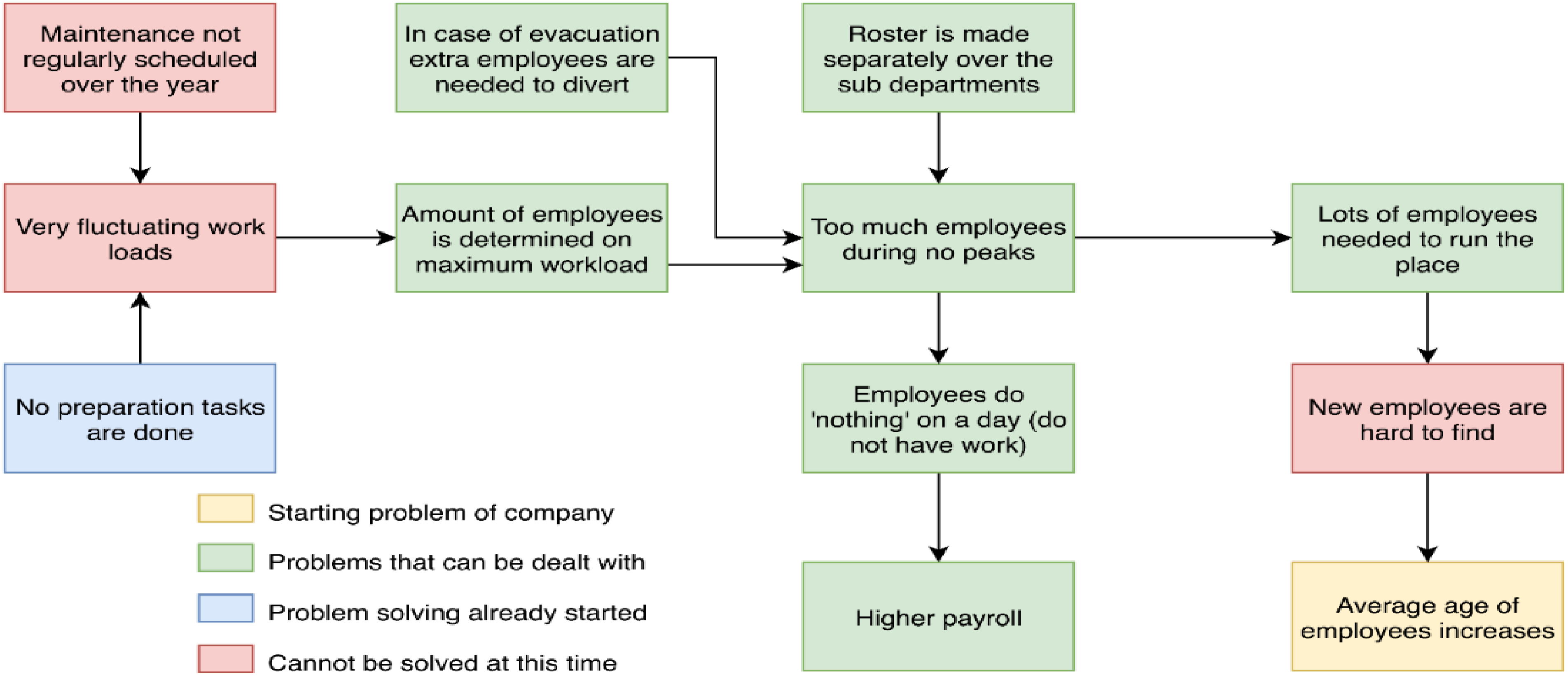

Figure 24 Problem cluster

Appendix C: Data analysis

This appendix gives the details of the steps made to come to a distribution. The steps are explained in *2.4 Simulation model*

"Simulation studies are computer experiments that involve creating data by pseudorandom sampling. The key strength of simulation studies is the ability to understand the behaviour of statistical methods because some 'truth' (usually some parameters of interest) is known from the process of generating the data." . This quote out of the work of Morris, White and Crowther describes very well what a simulation study is. In short a simulation model is a recreation of the reality to be able to value different chances in comparison to the reality.

The first step of making a simulation model is creating a conceptual model. Figure 4 gives a schematic overview of what the steps are during conceptual modelling. The conceptual model consists of the input for the model, activity diagrams, assumptions, simplifications, restrictions and output of the model. In fact the conceptual model is the complete description of what to use and what to calculate during simulation.

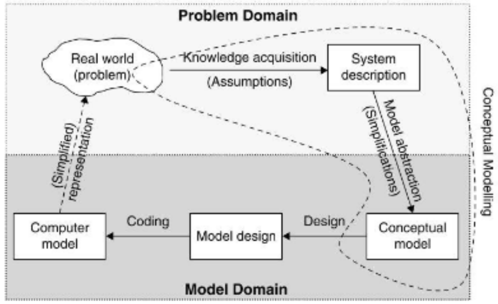

Figure 3: Build up simulation model (Robinson, 2014)

The step after the conceptual model is the model design. This consist mostly of collecting data and finding distributions to fit this data. After finishing this step, all the flows of the jobs through the system are known together with the distribution for the duration and the number of jobs. This is all the data needed for the model and thus the coding can be done.

The coding can be done in many different software programs. Think of Simul8, AnyLogic, SimScale or Tecnomatix Plant Simulation (TPS). For this thesis TPS will be used, as this software is known to me.

Nature of simulation models and simulation output

Simulation models can be terminating or non-terminating simulations . Terminating simulations have a natural endpoint. This can be the closing time of a shop or the end of the time-period of the investigation or the completion of a schedule. . Non-terminating simulations will never stop. "There is no reason for the simulation to stop other than the user interrupting the run.". The length of the simulation has to be determined by the user for these kind of simulations.

The department of OBI is a 24/7 active control centre. This means that the simulation should be a non-terminating simulation.

Non-terminating simulations often reach a steady-state . Steady-state means that the output of the KPI's will vary according to some sort of a distribution, called the steady-state distribution. This steady-state can be two forms. It can be a steady state point/line, meaning that after reaching this steady-state the output will vary according to the point/line. Another option is the steady-state cycle. This means that there is a certain pattern that keeps repeating itself.

Warm-up length

For the validity of the output it is important to know the warm-up length. This can be determined by making use of the Marginal Standard Error Rule (MSER) or a graphical method (same as replications graphical method). "MSER has the aim to minimise the width of the confidence interval about the mean of the simulation output data." .

$$MSER(d) = \frac{1}{(m-d)\cdot 2}\sum_{i=d+1}^{m}\left(Y_i - \bar{Y}(m,d)\right)^2$$

d = the proposed warm-up period
m = the number of observations in the time-series of output data
$\bar{Y}(m,d)$ = the mean of the observations from Y_{d+1} to Y_m

For every *d*, the MSER will be calculated. This means that for every *d,* the difference between the current mean and the mean of all the data coming up later is determined. The *d* which has the lowest MSER will be the warm-up period. Reason for this is that the difference of the current KPI value and the value in the future of the simulation is the lowest.
There is one constraint to the determination of the warm-up period. If the warm-up period is more than half of the run length used to determine the MSER, this value is not accountable and the MSER has to be determined out of a larger data sample.

Replications

Not only the warm-up period is important to know, also the length of the run and the amount of replications of the run are important. Running the simulation too short can cause that the data is too little to make proper analysis out of. However, running the simulation very long increases the run-time. The same counts for the number of replications. The more replications, the more data there is, but the runtime increases as well. Therefore, for both terms, the distinction has to be made between runtime and the amount of data.

"A replication is a run of a simulation model that uses specific streams of random numbers." . By changing the streams of random numbers, more randomness is inserted in the model, resulting in a more accurate output.

Robinson (2014) describes three approaches to come up with the amount of replications, namely: a rule of thumb, a graphical method and a confidence interval method.

A rule of thumb

Robinson (2014) states that according to Law and McComas (1990) at least three to five replications are needed. This method does not look into the output of a model, but is just an advice that could be followed.

Graphical method

The graphical method uses the cumulative mean of multiple replications to determine the amount of replications needed. The first step of this method is to make a lot of replications of a simulation model (around 20). For every of these replications, the runtime is the same and the warm-up period is used. Every replication gives a KPI value. A graph can be made with all cumulative means, one mean for every amount of replications. The point at which the graph is nearly flat, is the amount of replications needed.

Confidence interval method

This method is an addition to the graphical method. Instead of plotting the cumulative means, a confidence interval is made. The goal of this method is to achieve a deviation percentage lower than 5%. In the book of Robinson (2014), this method is explained in detail. This method uses formulas to determine the actual deviation for every number of replications.

Because of the 10 week time constraint and because of the little difference between the confidence interval method and the graphical method, it is chosen to use the graphical method for both the replications as well as the warm-up length. This method is a lot quicker to determine and almost as accurate as the confidence interval method/MSER. Because of the decision to use the graphical method, the confidence interval method is only shortly described, as it is of no usage for this research.

Run length

After determining the number of replications, the run-length should be determined. There is almost no information on how to calculate the run-length. However, according to Banks et al. (2009), the run-length should at least be 10 times the warm-up period. Therefore this will be used in the thesis.

.

Performing switch sequences

The duration of executing a switch sequence will be determined per station. The dataset used is historic data. It is retrieved from the logbooks which consist of all the actions done at the department.

This section will be divided into the two stations, so North and South.

North

Step 1: Graphical display of the dataset

Figure 25: Duration of sequences North divides the dataset in a couple of bins. The density of the bin is put into a graph. What can be noticed is that the skewness is aimed to the left. Furthermore, the dataset consist of some very long durations. In the cumulative distribution this can be noticed. There are 2 points around the 700 seconds. As these points are so far away from the other points, these points will not be taken into account during the fitting of a distribution.

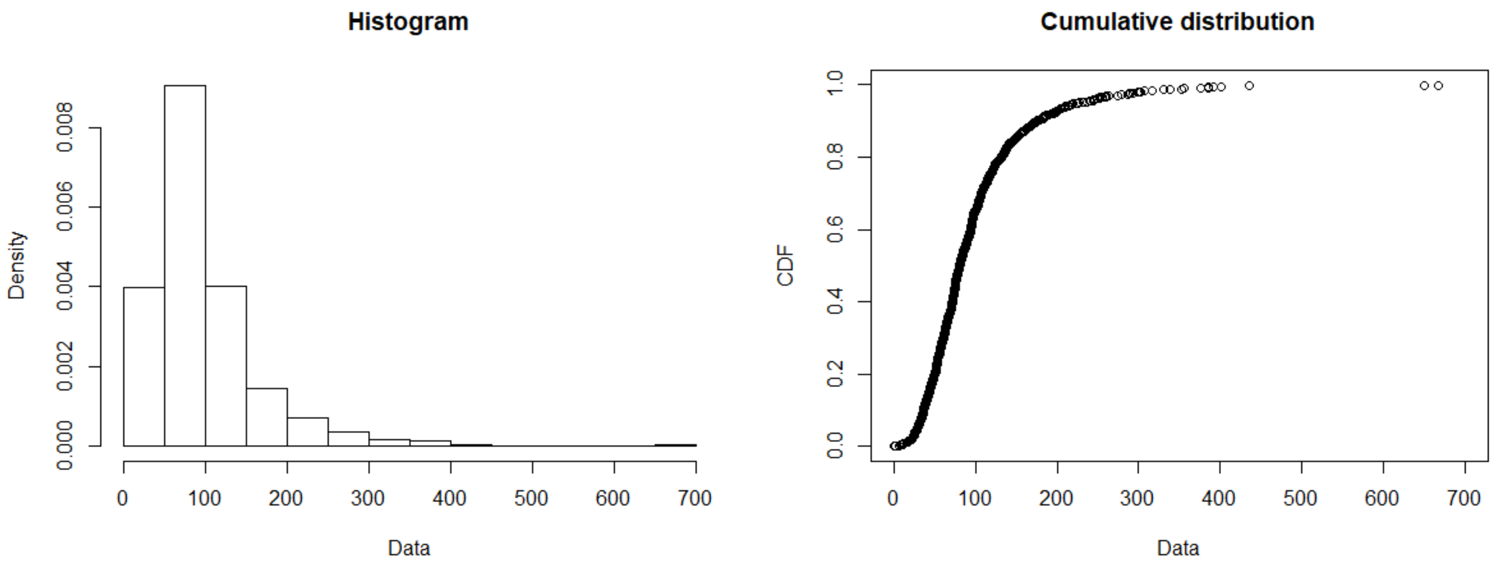

Figure 25: Duration of sequences North

Step 2: Characterization of the dataset

Figure 26: C&F duration sequences North is the Cullen and Frey graph. This graph calculates the kurtosis and the skewness. By comparing these values with the average skewness and kurtosis of different distributions, it can be seen which distributions lie close to the dataset. In this case, the gamma, Weibull and lognormal distributions are close to the dataset.

```
summary statistics
------
min:  0.2   max:  667
median:  80.9
mean:  98.29921
estimated sd:  70.71739
```

```
estimated skewness:  2.953368
estimated kurtosis:  18.40628
```

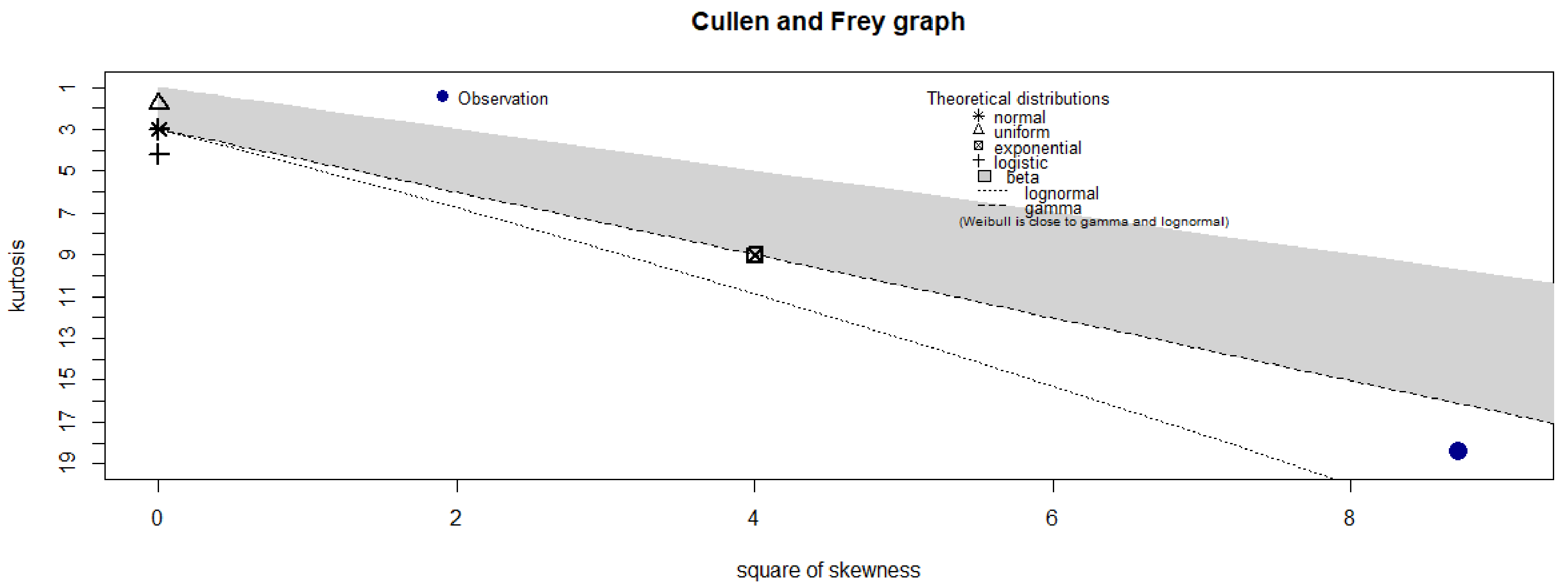

Figure 26: C&F duration sequences North

Step 3: Fitting of a distribution to the dataset

During this step, the distributions said in step 2 are made into graphs. In this step only the distribution that is correct is shown. This is just an indication how the distribution fits the data. Furthermore, the summary of the distribution parameters is given.

```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
          estimate Std. Error
meanlog 4.3858964 0.02229942
sdlog   0.6618834 0.01576791
Loglikelihood:  -4750.501   AIC:  9505.002   BIC:  9514.564
Correlation matrix:
```

```
          meanlog sdlog
meanlog         1     0
sdlog           0     1
```

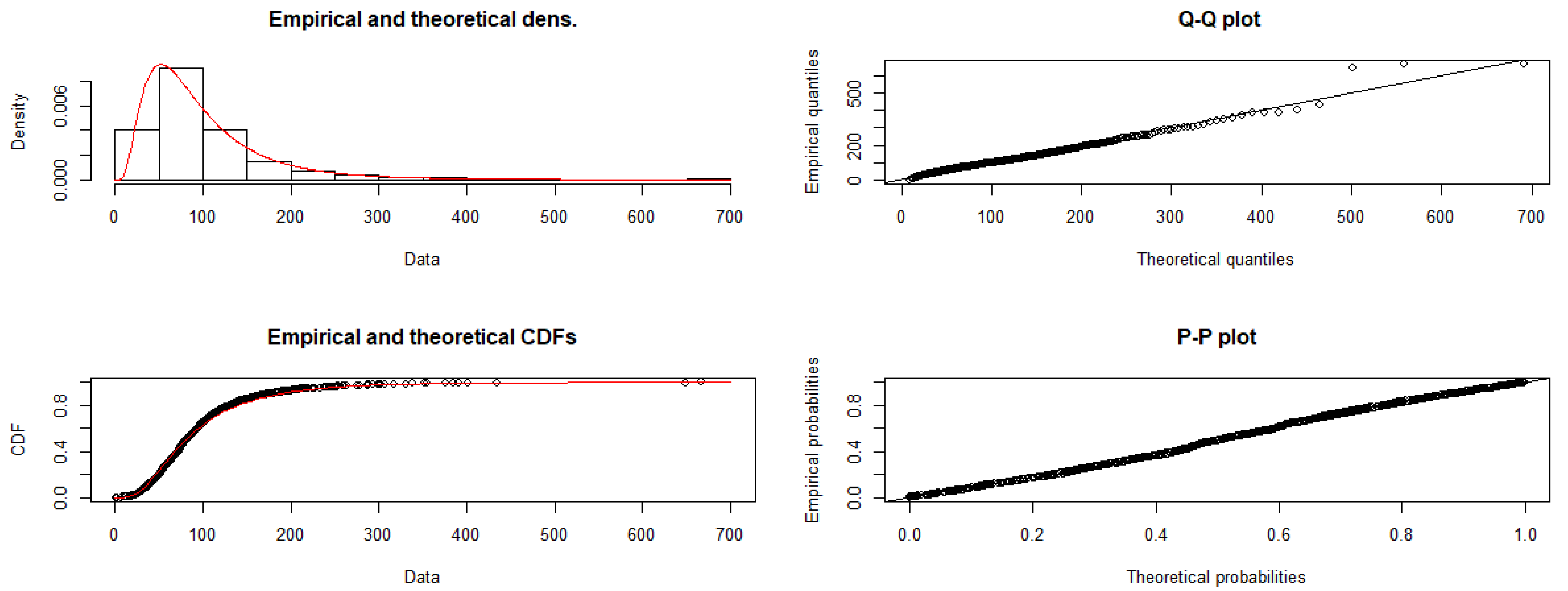

Figure 27: Fitting Lognormal distribution North

```
        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.045028, p-value = 0.05617
alternative hypothesis: two-sided
```

What can be seen in the data above, is that the lognormal distribution fits the data and has a p-value which is larger than 0.05. This means that the H_0 hypothesis cannot be rejected and thus that there is no evidence that the dataset does not follow a lognormal distribution. Therefore, the lognormal distribution will be taken as the representative distribution.

Step 4: Simulation of the uncertainty of the distribution to the dataset

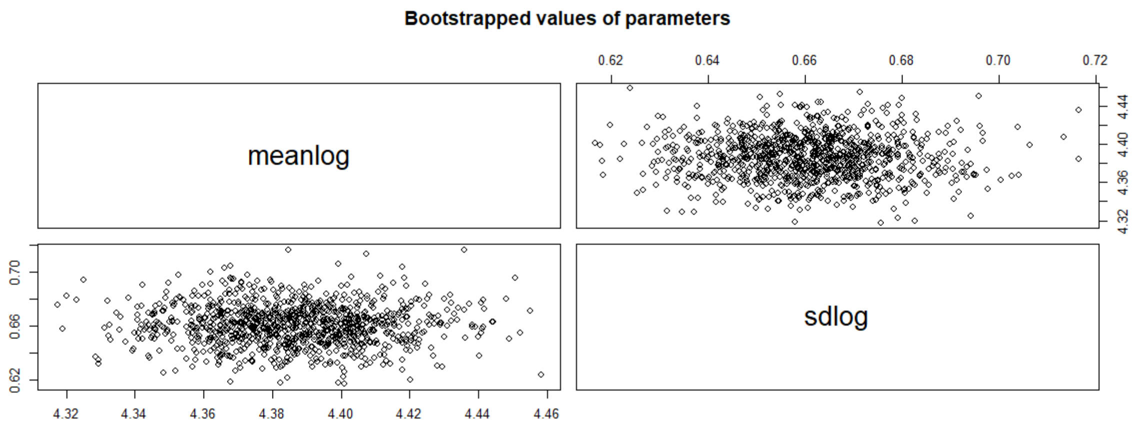

Figure 28: Bootstrapping duration of sequences North

```
Parametric bootstrap medians and 95% percentile CI
           Median      2.5%     97.5%
meanlog 4.3859028 4.3410615 4.4331237
sdlog   0.6607082 0.6310385 0.6921763
```

By bootstrapping the parameters, it comes up with an insight in the confidence interval of the data. By using the median and the 97.5 % percentile as boundaries, the parameter can change between those, to make sure that it is most representative for the real world.

South

Step 1: Graphical display of the dataset

Figure 29: Duration of sequences South divides the dataset in a couple of bins. The density of the bin is put into a graph. What can be noticed is that the skewness is aimed to the left. Furthermore, the dataset consist of some very long durations. In the cumulative distribution this can be noticed. There is 1 point around the 500 seconds. As this point is so far away from the other points, this point will not be taken into account during the fitting of a distribution.

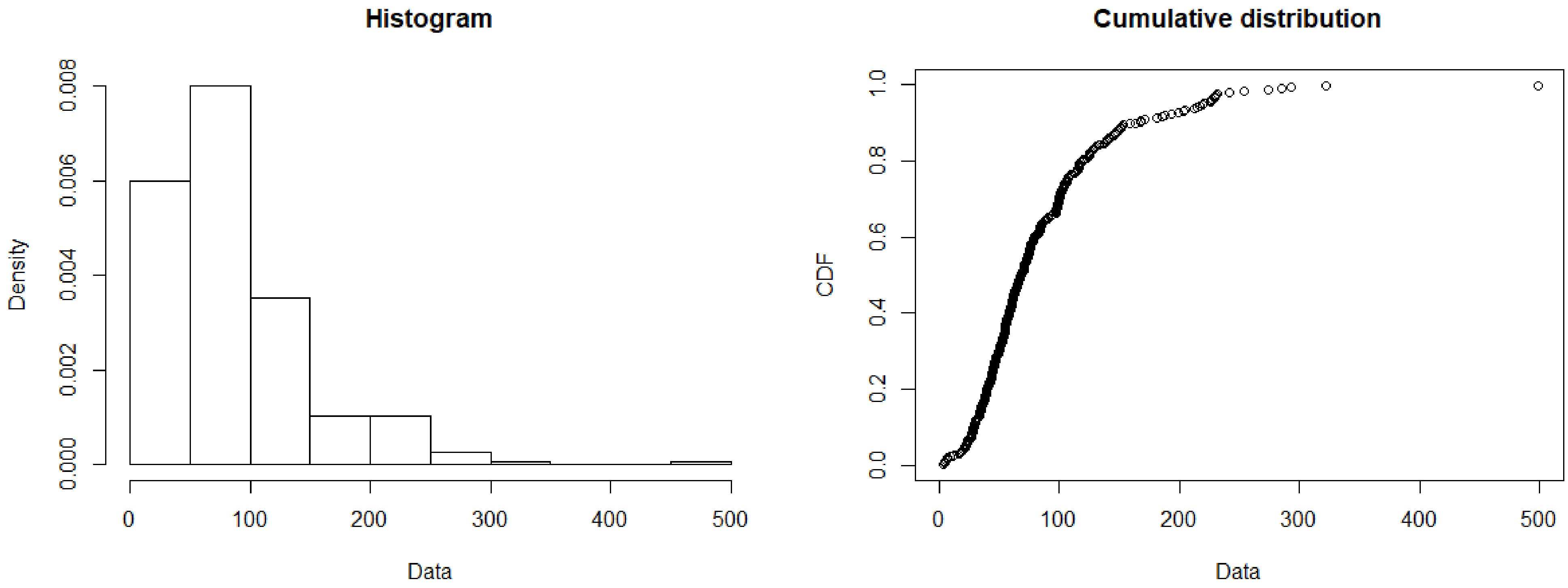

Figure 29: Duration of sequences South

Step 2: Characterization of the dataset

Figure 26: C&F duration sequences North is the Cullen and Frey graph. This graph calculates the kurtosis and the skewness. By comparing these values with the average skewness and kurtosis of different distributions, it can be seen which distributions lie close to the dataset. In this case, the gamma, Weibull, exponential and lognormal distributions are close to the dataset.

```
summary statistics
------
min:  3.8   max:  498.899
median:  68.6495
mean:  85.9262
estimated sd:  62.40275
estimated skewness:  2.075124
estimated kurtosis:  10.31768
```

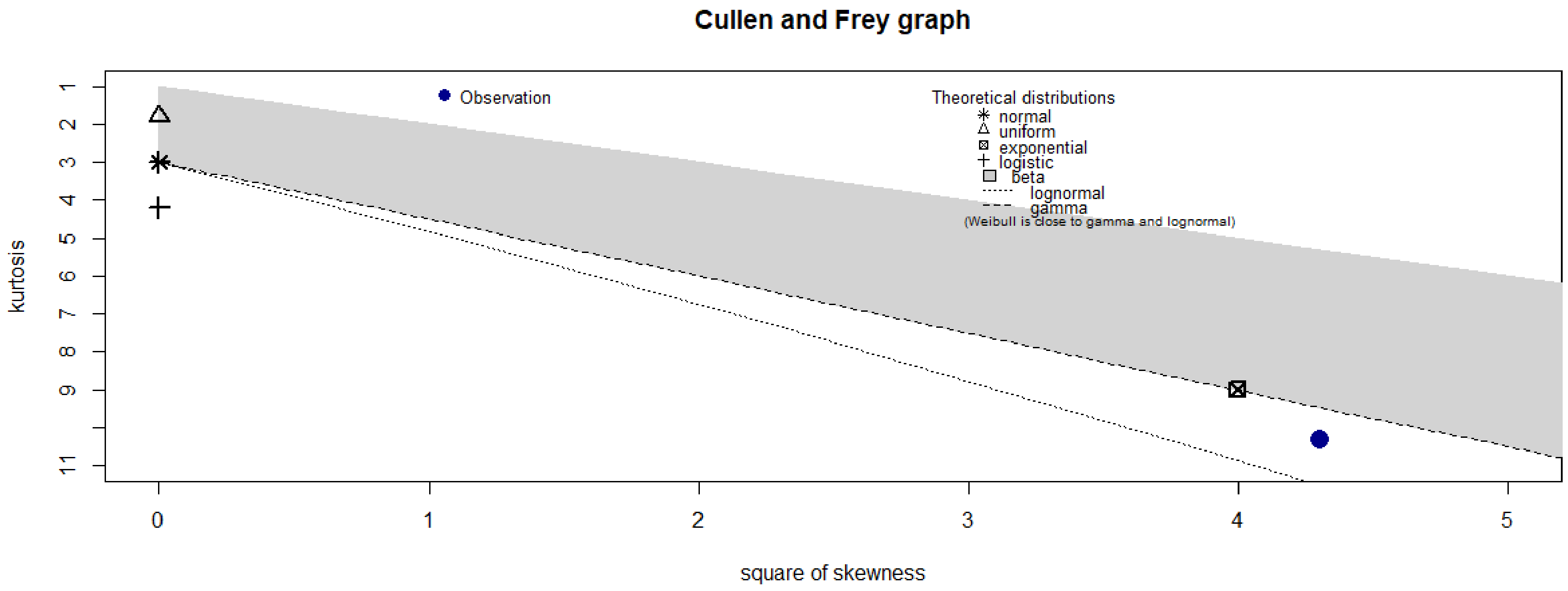

Figure 30: C&F duration sequences South

Step 3: Fitting of a distribution to the dataset

During this step, the distributions said in step 2 are made into graphs. In this step only the distribution that is correct is shown. This is just an indication how the distribution fits the data. Furthermore, the summary of the distribution parameters is given.

```
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters :
       estimate Std. Error
shape  1.501632 0.06371786
scale 95.746384 3.93482092
Loglikelihood:  -1567.15   AIC:  3138.3   BIC:  3145.667
Correlation matrix:
          shape     scale
shape 1.0000000 0.3271723
scale 0.3271723 1.0000000
```

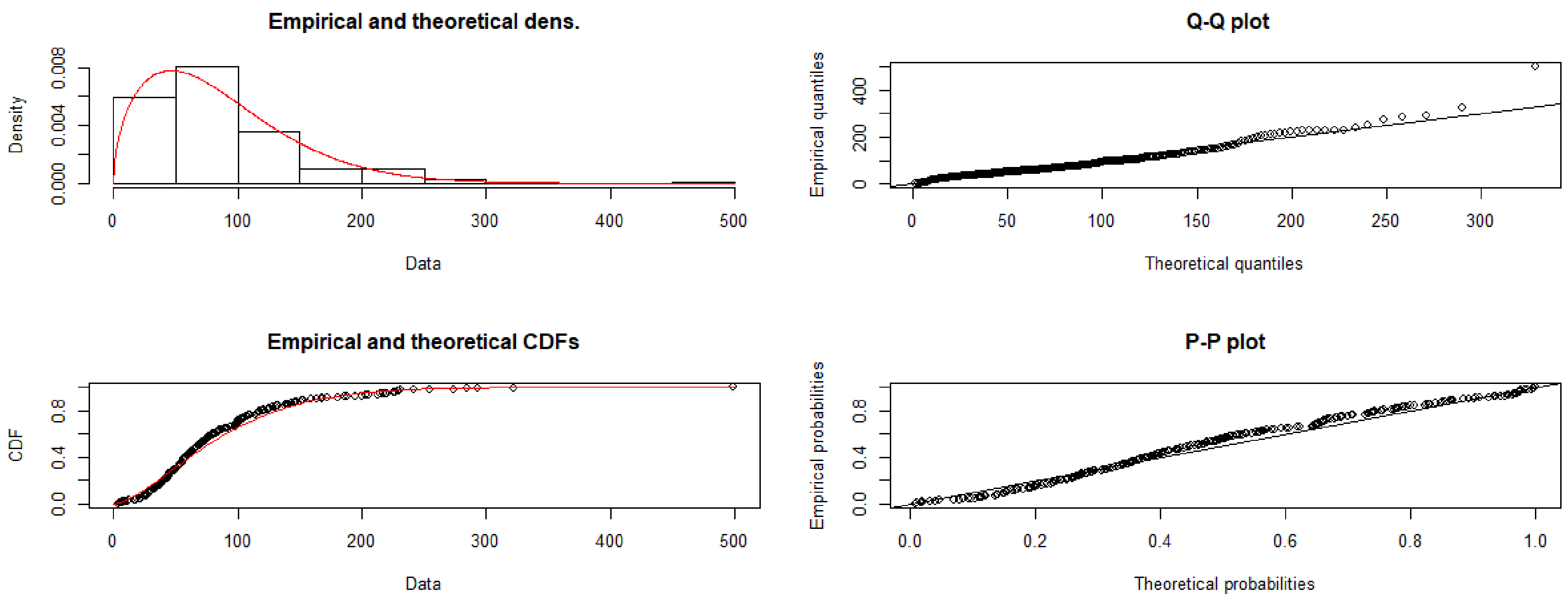

Figure 31: Fitting Weibull distribution South

For the South 1500V station, the duration of the calls can be represented by a Weibull distribution. This distribution fits the dataset the best and can represent the dataset according to the p-value out of the Kolmogorov-Smirnov test.

```
        One-sample Kolmogorov-Smirnov test

data:  y
D = 0.069853, p-value = 0.1135
alternative hypothesis: two-sided
```

Step 4: Simulation of the uncertainty of the distribution to the dataset

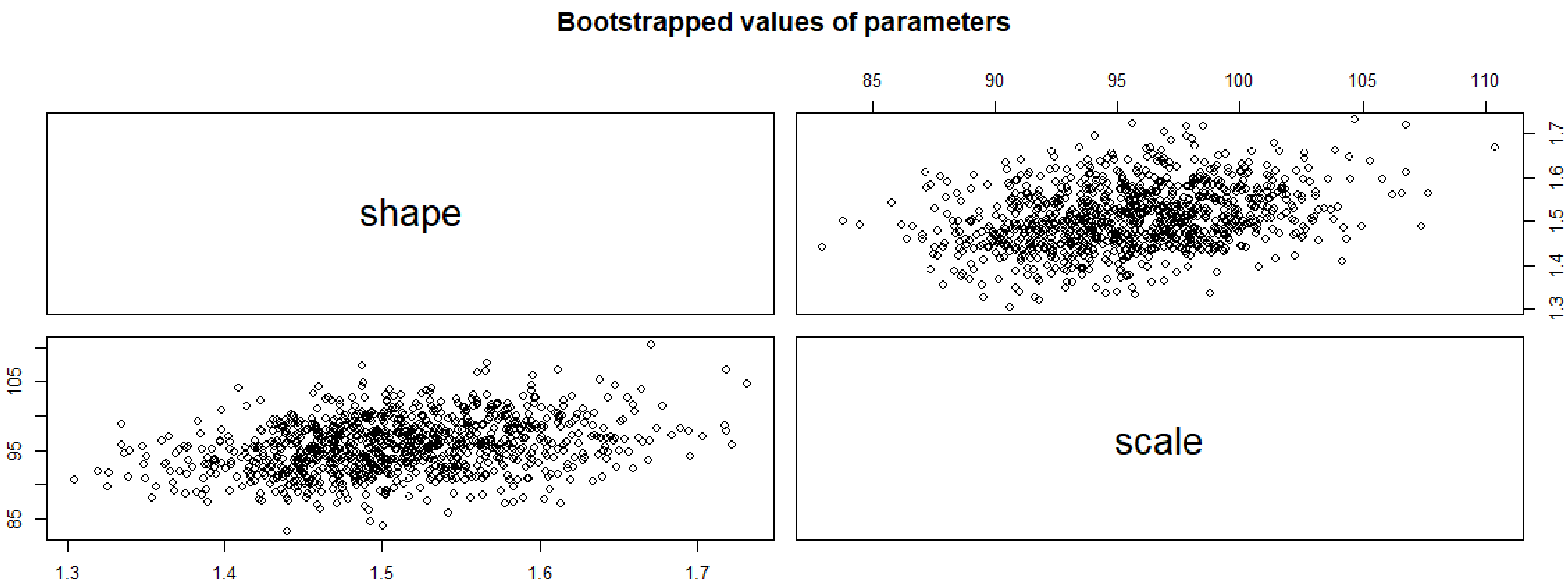

Figure 32: Bootstrapping duration of sequences South

By bootstrapping the parameters, it comes up with an insight in the confidence interval of the data. By using the median and the 97.5 % percentile as boundaries, the parameter can change between those, to make sure that it is most representative for the real world.

Appendix D: Solution distribution fit

Combi			
Time	Distribu	Par1	Par2
00:00	Gamma	1,343	0,019
01:00	Gamma	1,412	0,017
02:00	Gamma	1,428	0,019
03:00	Gamma	1,338	0,019
04:00	Gamma	1,627	0,024
05:00	Gamma	1,621	0,021
06:00	Gamma	1,695	0,023
07:00	Lognorma	3,927	1,047
08:00	Gamma	1,153	0,802
09:00	Gamma	1,010	0,012
10:00	Lognorma	3,758	1,063
11:00	Gamma	0,981	0,012
12:00	Lognorma	3,828	1,075
13:00	Lognorma	1,872	1,059
14:00	Gamma	1,127	0,015
15:00	Lognorma	3,813	1,177
16:00	Lognorma	3,736	1,233
17:00	Gamma	1,104	0,122
18:00	Gamma	1,350	0,016
19:00	Gamma	1,480	0,017
20:00	Gamma	0,910	0,008
21:00	Gamma	0,930	0,010
22:00	Gamma	1,038	0,011
23:00	Gamma	1,233	0,018

North			
Time	Distribution	Par1	Par2
00:00	Gamma	1,347	0,019
01:00	Gamma	1,347	0,019
02:00	Gamma	1,596	0,023
03:00	Gamma	1,198	0,018
04:00	Gamma	1,198	0,018
05:00	Weibull	1,199	79,438
06:00	Gamma	1,948	0,028
07:00	Lognormal	3,978	0,895
08:00	Lognormal	3,810	1,005
09:00	Lognormal	3,810	1,005
10:00	Lognormal	3,810	1,005
11:00	Lognormal	3,803	0,961
12:00	Lognormal	3,803	0,961
13:00	Lognormal	3,803	0,961
14:00	Lognormal	3,737	1,067
15:00	Lognormal	3,737	1,067
16:00	Lognormal	3,757	1,163
17:00	Lognormal	3,946	1,198
18:00	Lognormal	3,965	0,915
19:00	Lognormal	3,965	0,915
20:00	Lognormal	4,222	1,182
21:00	Lognormal	4,257	1,089
22:00	Lognormal	3,895	1,256
23:00	Lognormal	3,895	1,256

South			
Time	Distribution	Par1	Par2
00:00	Gamma	1,338	0,019
01:00	Gamma	1,417	0,016
02:00	Gamma	1,275	0,015
03:00	Gamma	1,499	0,020
04:00	Gamma	1,499	0,020
05:00	Gamma	1,710	0,022
06:00	Lognormal	4,025	0,887
07:00	Gamma	1,076	0,013
08:00	Gamma	1,076	0,013
09:00	Gamma	1,076	0,013
10:00	Lognormal	3,915	1,073
11:00	Gamma	0,905	0,011
12:00	Gamma	0,905	0,011
13:00	Lognormal	3,915	1,073
14:00	Gamma	0,957	0,013
15:00	Gamma	1,221	0,014
16:00	Gamma	0,957	0,013
17:00	Gamma	1,221	0,014
18:00	Gamma	1,456	0,017
19:00	Gamma	1,873	0,019
20:00	Gamma	1,036	0,010
21:00	Gamma	0,834	0,010
22:00	Gamma	1,132	0,012
23:00	Gamma	1,563	0,021

Figure 33: Distributions for the stations

Appendix E: Solutions simulation model

North (Current)		North (New)		Combi (Current)		Combi (New)	
Replications	Waiting time	Replications	Waiting time	Replications	Waiting time	Replications	Waiting time
1	218,4935	1	563,9923	1	1630,7342	1	2748,2371
2	221,8756	2	555,14	2	1624,231	2	2751,4785
3	216,95545	3	559,82	3	1626,458	3	2750,147
Average	219,1081833	Average	559,6507667	Average	1627,141067	Average	2749,9542

Figure 34: Solutions simulations model

North (advice)	
Replications	Waiting time
1	393,5421
2	397,7156
3	396,8945
Average	396,0507

Appendix F: Distributions and variables

This appendix shows the variables used in the simulation and mathematical model. The duration of the calls can be found in appendix D.

Table 6: Number of sequences per day

Number of sequences			
	North	**South**	**Combi**
Monday	5,361023	2,816712	7,384286
Tuesday	4,934456	1,659744	6,107339
Wednesday	5,440691	2,307756	7,27783
Thursday	4,019015	2,50609	5,856513
Friday	3,230527	1,585518	4,277855
Saturday	2,018716	1,451949	2,75668
Sunday	3,504262	2,593946	5,176487

Table 7: Duration of performing a sequence

Duration of performing a sequence	
North	150,546
South	122,499
Combi	137,657

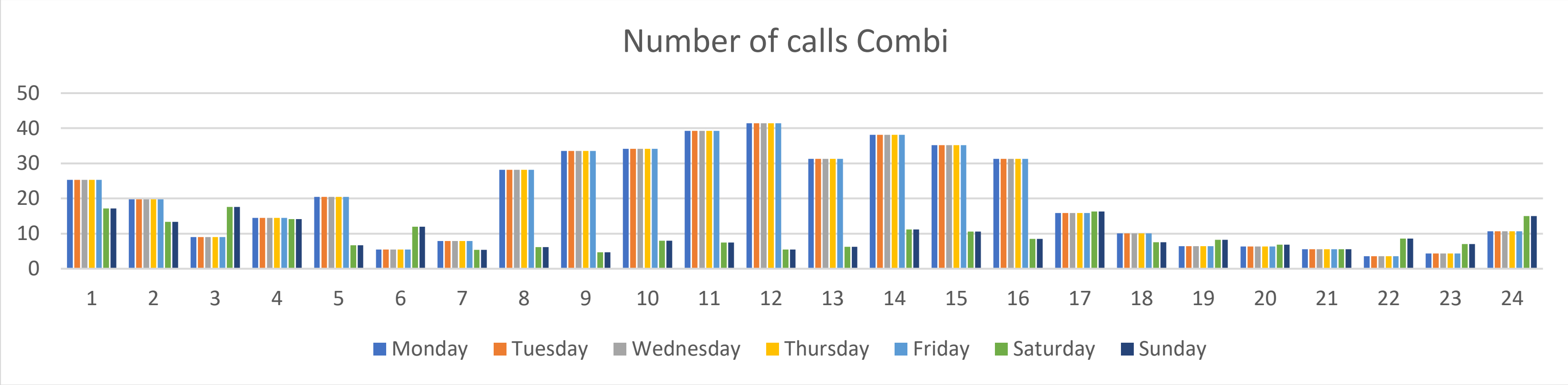

Figure 35: Number of calls Combi station

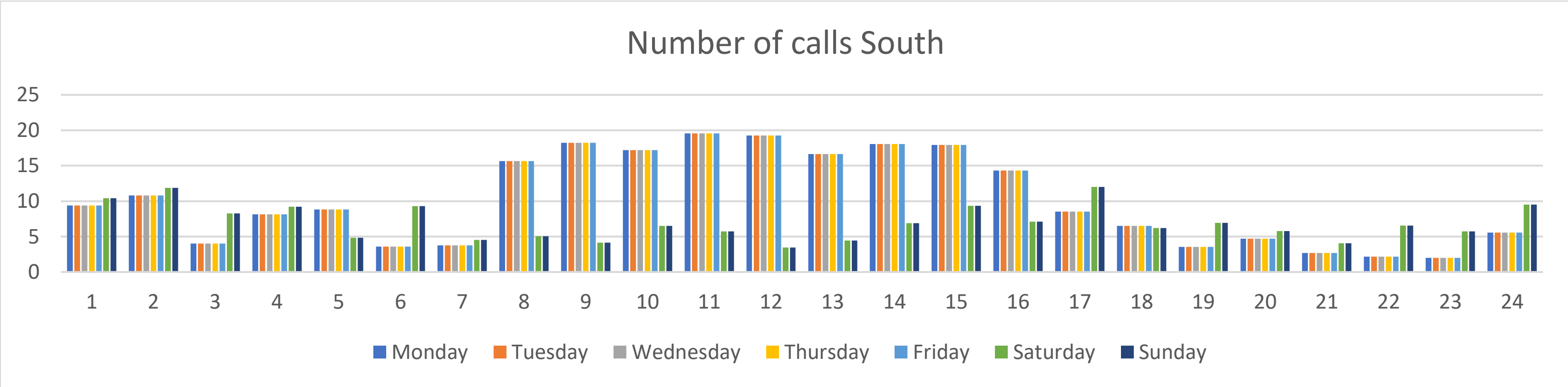

Figure 36: Number of calls South station

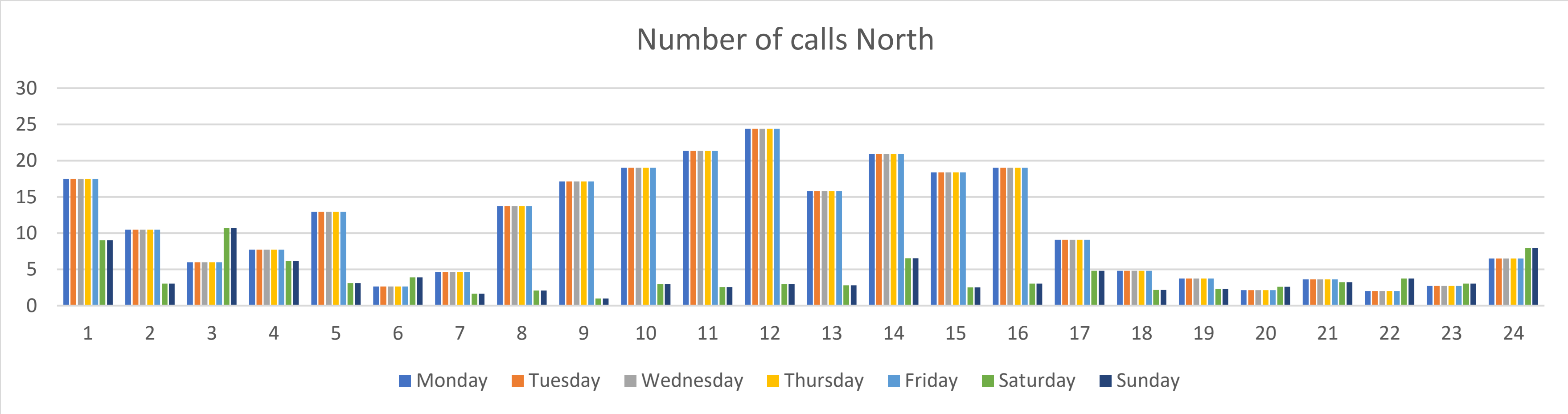

Figure 37: Number of calls North station